# Intégration de données ouvertes avec WebSmatch

Patrick Valduriez

Remi Coletta, Emmanuel Castanier,
Christian Frisch, DuyHoa Ngo, Zohra Bellahsene





# Motivations

**Context:** open data in France

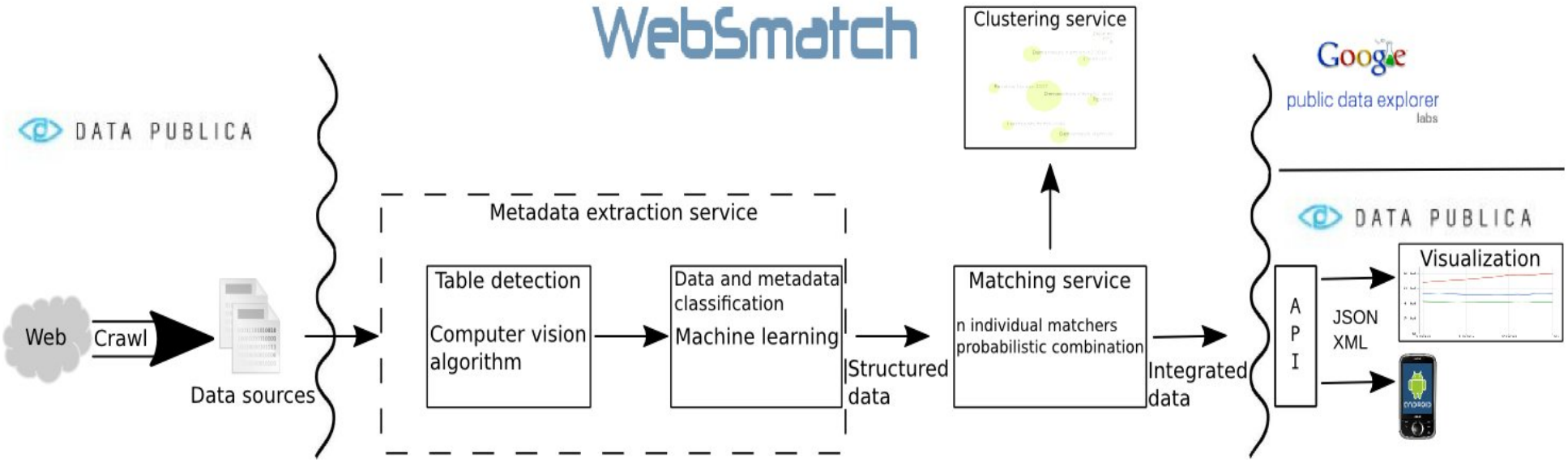
**Problems**

- High number of data sources
- Heterogeneous formats
- Poorly structured

**Example (Data Publica):** the web crawl for french open data sources found 148509 Excel files and only 369 RDF files

**Needs:** integrate data sources to yield high-value information

# WebSmatch - Data Publica workflow



# Example of input data source

**URL :** http://www.data-publica.com/publication/4736

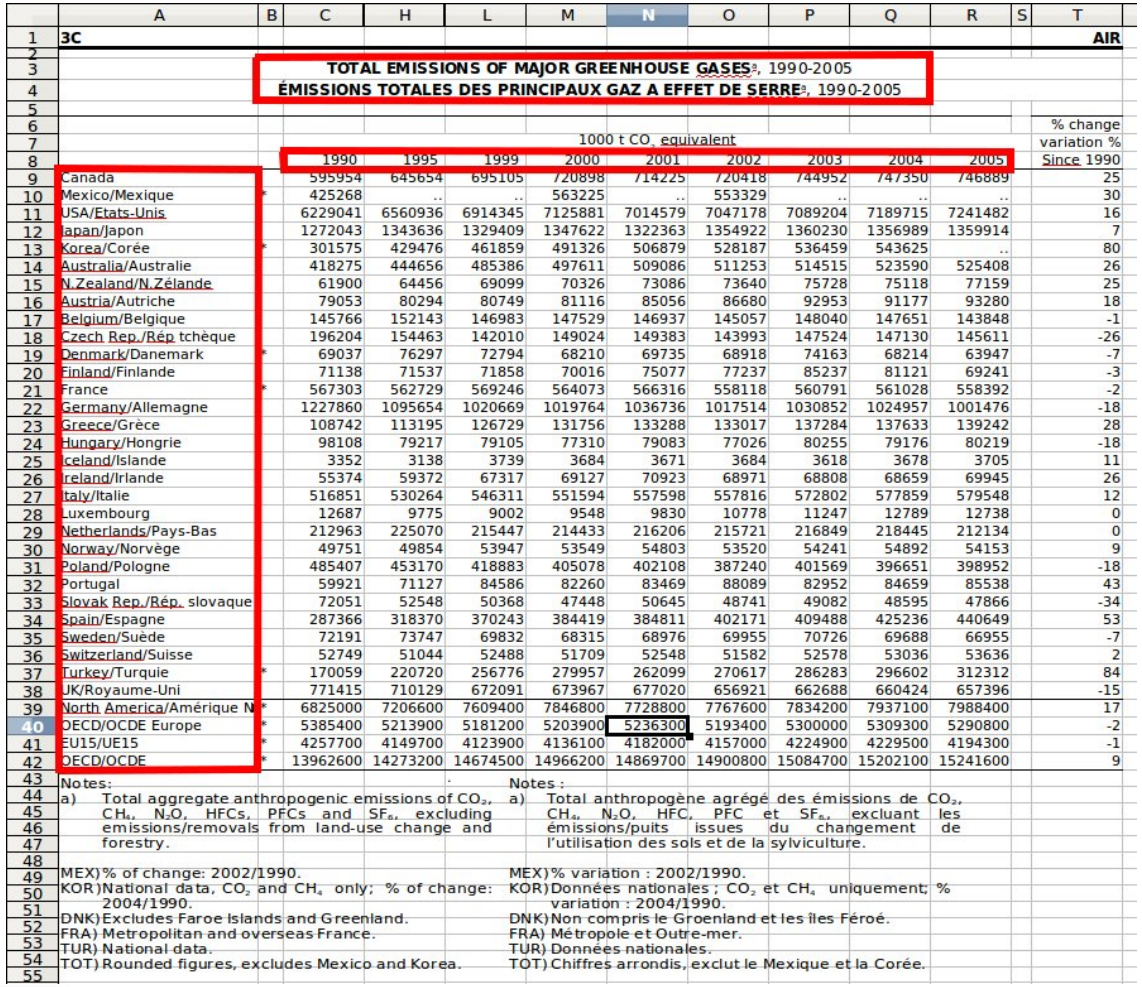
| | A | B | C | H | L | M | N | O | P | Q | R | S | T |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 3C | | | | | | | | | | | | AIR |
| 2 | | | | | | | | | | | | | |
| 3 | TOTAL EMISSIONS OF MAJOR GREENHOUSE GASES[a], 1990-2005 | | | | | | | | | | | | |
| 4 | ÉMISSIONS TOTALES DES PRINCIPAUX GAZ A EFFET DE SERRE[a], 1990-2005 | | | | | | | | | | | | |
| 5 | | | | | | | | | | | | | |
| 6 | | | | | | | | | | | | | % change |
| 7 | | | 1000 t $CO_2$ equivalent | | | | | | | | | | variation % |
| 8 | | | 1990 | 1995 | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | | Since 1990 |
| 9 | Canada | | 595954 | 645654 | 695105 | 720898 | 714225 | 720418 | 744952 | 747350 | 746889 | | 25 |
| 10 | Mexico/Mexique | * | 425268 | .. | .. | 563225 | .. | 553329 | .. | .. | .. | | 30 |
| 11 | USA/Etats-Unis | | 6229041 | 6560936 | 6914345 | 7125881 | 7014579 | 7047178 | 7089204 | 7189715 | 7241482 | | 16 |
| 12 | Japan/Japon | | 1272043 | 1343636 | 1329409 | 1347622 | 1322363 | 1354922 | 1360230 | 1356989 | 1359914 | | 7 |
| 13 | Korea/Corée | * | 301575 | 429476 | 461859 | 491326 | 506879 | 528187 | 536459 | 543625 | .. | | 80 |
| 14 | Australia/Australie | | 418275 | 444656 | 485386 | 497611 | 509086 | 511253 | 514515 | 523590 | 525408 | | 26 |
| 15 | N.Zealand/N.Zélande | | 61900 | 64456 | 69099 | 70326 | 73086 | 73640 | 75728 | 75118 | 77159 | | 25 |
| 16 | Austria/Autriche | | 79053 | 80294 | 80749 | 81116 | 85056 | 86680 | 92953 | 91177 | 93280 | | 18 |
| 17 | Belgium/Belgique | | 145766 | 152143 | 146983 | 147529 | 146937 | 145057 | 148040 | 147651 | 143848 | | -1 |
| 18 | Czech Rep./Rép tchèque | | 196204 | 154463 | 142010 | 149024 | 149383 | 143993 | 147524 | 147130 | 145611 | | -26 |
| 19 | Denmark/Danemark | * | 69037 | 76297 | 72794 | 68210 | 69735 | 68918 | 74163 | 68214 | 63947 | | -7 |
| 20 | Finland/Finlande | | 71138 | 71537 | 71858 | 70016 | 75077 | 77237 | 85237 | 81121 | 69241 | | -3 |
| 21 | France | * | 567303 | 562729 | 569246 | 564073 | 566316 | 558118 | 560791 | 561028 | 558392 | | -2 |
| 22 | Germany/Allemagne | | 1227860 | 1095654 | 1020669 | 1019764 | 1036736 | 1017514 | 1030852 | 1024957 | 1001476 | | -18 |
| 23 | Greece/Grèce | | 108742 | 113195 | 126729 | 131756 | 133288 | 133017 | 137284 | 137633 | 139242 | | 28 |
| 24 | Hungary/Hongrie | | 98108 | 79217 | 79105 | 77310 | 79083 | 77026 | 80255 | 79176 | 80219 | | -18 |
| 25 | Iceland/Islande | | 3352 | 3138 | 3739 | 3684 | 3671 | 3684 | 3618 | 3678 | 3705 | | 11 |
| 26 | Ireland/Irlande | | 55374 | 59372 | 67317 | 69127 | 70923 | 68971 | 68808 | 68659 | 69945 | | 26 |
| 27 | Italy/Italie | | 516851 | 530264 | 546311 | 551594 | 557598 | 557816 | 572802 | 577859 | 579548 | | 12 |
| 28 | Luxembourg | | 12687 | 9775 | 9002 | 9548 | 9830 | 10778 | 11247 | 12789 | 12738 | | 0 |
| 29 | Netherlands/Pays-Bas | | 212963 | 225070 | 215447 | 214433 | 216206 | 215721 | 216849 | 218445 | 212134 | | 0 |
| 30 | Norway/Norvège | | 49751 | 49854 | 53947 | 53549 | 54803 | 53520 | 54241 | 54892 | 54153 | | 9 |
| 31 | Poland/Pologne | | 485407 | 453170 | 418883 | 405078 | 402108 | 387240 | 401569 | 396651 | 398952 | | -18 |
| 32 | Portugal | | 59921 | 71127 | 84586 | 82260 | 83469 | 88089 | 82952 | 84659 | 85538 | | 43 |
| 33 | Slovak Rep./Rép. slovaque | | 72051 | 52548 | 50368 | 47448 | 50645 | 48741 | 49082 | 48595 | 47866 | | -34 |
| 34 | Spain/Espagne | | 287366 | 318370 | 370243 | 384419 | 384811 | 402171 | 409488 | 425236 | 440649 | | 53 |
| 35 | Sweden/Suède | | 72191 | 73747 | 69832 | 68315 | 68976 | 69955 | 70726 | 69688 | 66955 | | -7 |
| 36 | Switzerland/Suisse | | 52749 | 51044 | 52488 | 51709 | 52548 | 51582 | 52578 | 53036 | 53636 | | 2 |
| 37 | Turkey/Turquie | * | 170059 | 220720 | 256776 | 279957 | 262099 | 270617 | 286283 | 296602 | 312312 | | 84 |
| 38 | UK/Royaume-Uni | | 771415 | 710129 | 672091 | 673967 | 677020 | 656921 | 662688 | 660424 | 657396 | | -15 |
| 39 | North America/Amérique N | * | 6825000 | 7206600 | 7609400 | 7846800 | 7728800 | 7767600 | 7834200 | 7937100 | 7988400 | | 17 |
| 40 | OECD/OCDE Europe | * | 5385400 | 5213900 | 5181200 | 5203900 | 5236300 | 5193400 | 5300000 | 5309300 | 5290800 | | -2 |
| 41 | EU15/UE15 | * | 4257700 | 4149700 | 4123900 | 4136100 | 4182000 | 4157000 | 4224900 | 4229500 | 4194300 | | -1 |
| 42 | OECD/OCDE | * | 13962600 | 14273200 | 14674500 | 14966200 | 14869700 | 14900800 | 15084700 | 15202100 | 15241600 | | 9 |
| 43 | | | | | | | | | | | | | |
| 44 | Notes: | | | | | Notes : | | | | | | | |
| 45 | a) Total aggregate anthropogenic emissions of $CO_2$, | | | | | a) Total anthropogène agrégé des émissions de $CO_2$, | | | | | | | |
| 46 | $CH_4$, $N_2O$, HFCs, PFCs and $SF_6$, excluding | | | | | $CH_4$, $N_2O$, HFC, PFC et $SF_6$, excluant les | | | | | | | |
| 47 | emissions/removals from land-use change and | | | | | émissions/puits issues du changement de | | | | | | | |
| 48 | forestry. | | | | | l'utilisation des sols et de la sylviculture. | | | | | | | |
| 49 | MEX) % of change: 2002/1990. | | | | | MEX) % variation : 2002/1990. | | | | | | | |
| 50 | KOR) National data, $CO_2$ and $CH_4$ only; % of change: | | | | | KOR) Données nationales ; $CO_2$ et $CH_4$ uniquement; % | | | | | | | |
| 51 | 2004/1990. | | | | | variation : 2004/1990. | | | | | | | |
| 52 | DNK) Excludes Faroe Islands and Greenland. | | | | | DNK) Non compris le Groenland et les îles Féroé. | | | | | | | |
| 53 | FRA) Metropolitan and overseas France. | | | | | FRA) Métropole et Outre-mer. | | | | | | | |
| 54 | TUR) National data. | | | | | TUR) Données nationales. | | | | | | | |
| 55 | TOT) Rounded figures, excludes Mexico and Korea. | | | | | TOT) Chiffres arrondis, exclut le Mexique et la Corée. | | | | | | | |

**Problem :** where are the data and metadata? incomplete lines, unnamed attributes

**Existing tools** such as OpenII or Google Refine work only on clean files

# Example of input data source

**URL :** http://www.data-publica.com/publication/4736

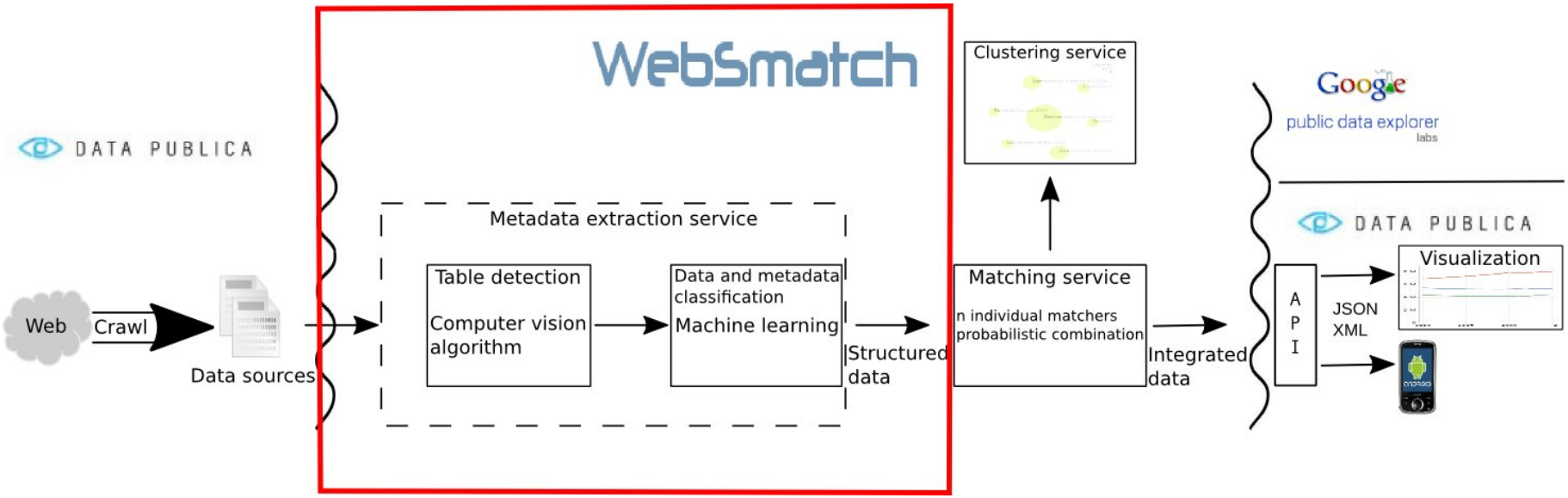


| | A | B | C | H | L | M | N | O | P | Q | R | S | T |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 3C | | | | | | | | | | | | AIR |
| 2 | | | | | | | | | | | | | |
| 3 | TOTAL EMISSIONS OF MAJOR GREENHOUSE GASES[a], 1990-2005 | | | | | | | | | | | | |
| 4 | ÉMISSIONS TOTALES DES PRINCIPAUX GAZ A EFFET DE SERRE[a], 1990-2005 | | | | | | | | | | | | |
| 5 | | | | | | | | | | | | | |
| 6 | | | | | | | | | | | | | % change |
| 7 | | | 1000 t $CO_2$ equivalent | | | | | | | | | | variation % |
| 8 | | | 1990 | 1995 | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | | Since 1990 |
| 9 | Canada | | 595954 | 645654 | 695105 | 720898 | 714225 | 720418 | 744952 | 747350 | 746889 | | 25 |
| 10 | Mexico/Mexique | * | 425268 | .. | .. | 563225 | .. | 553329 | .. | .. | .. | | 30 |
| 11 | USA/Etats-Unis | | 6229041 | 6560936 | 6914345 | 7125881 | 7014579 | 7047178 | 7089204 | 7189715 | 7241482 | | 16 |
| 12 | Japan/Japon | | 1272043 | 1343636 | 1329409 | 1347622 | 1322363 | 1354922 | 1360230 | 1356989 | 1359914 | | 7 |
| 13 | Korea/Corée | * | 301575 | 429476 | 461859 | 491326 | 506879 | 528187 | 536459 | 543625 | .. | | 80 |
| 14 | Australia/Australie | | 418275 | 444656 | 485386 | 497611 | 509086 | 511253 | 514515 | 523590 | 525408 | | 26 |
| 15 | N.Zealand/N.Zélande | | 61900 | 64456 | 69099 | 70326 | 73086 | 73640 | 75728 | 75118 | 77159 | | 25 |
| 16 | Austria/Autriche | | 79053 | 80294 | 80749 | 81116 | 85056 | 86680 | 92953 | 91177 | 93280 | | 18 |
| 17 | Belgium/Belgique | | 145766 | 152143 | 146983 | 147529 | 146937 | 145057 | 148040 | 147651 | 143848 | | -1 |
| 18 | Czech Rep./Rép tchèque | | 196204 | 154463 | 142010 | 149024 | 149383 | 143993 | 147524 | 147130 | 145611 | | -26 |
| 19 | Denmark/Danemark | * | 69037 | 76297 | 72794 | 68210 | 69735 | 68918 | 74163 | 68214 | 63947 | | -7 |
| 20 | Finland/Finlande | | 71138 | 71537 | 71858 | 70016 | 75077 | 77237 | 85237 | 81121 | 69241 | | -3 |
| 21 | France | * | 567303 | 562729 | 569246 | 564073 | 566316 | 558118 | 560791 | 561028 | 558392 | | -2 |
| 22 | Germany/Allemagne | | 1227860 | 1095654 | 1020669 | 1019764 | 1036736 | 1017514 | 1030852 | 1024957 | 1001476 | | -18 |
| 23 | Greece/Grèce | | 108742 | 113195 | 126729 | 131756 | 133288 | 133017 | 137284 | 137633 | 139242 | | 28 |
| 24 | Hungary/Hongrie | | 98108 | 79217 | 79105 | 77310 | 79083 | 77026 | 80255 | 79176 | 80219 | | -18 |
| 25 | Iceland/Islande | | 3352 | 3138 | 3739 | 3684 | 3671 | 3684 | 3618 | 3678 | 3705 | | 11 |
| 26 | Ireland/Irlande | | 55374 | 59372 | 67317 | 69127 | 70923 | 68971 | 68808 | 68659 | 69945 | | 26 |
| 27 | Italy/Italie | | 516851 | 530264 | 546311 | 551594 | 557598 | 557816 | 572802 | 577859 | 579548 | | 12 |
| 28 | Luxembourg | | 12687 | 9775 | 9002 | 9548 | 9830 | 10778 | 11247 | 12789 | 12738 | | 0 |
| 29 | Netherlands/Pays-Bas | | 212963 | 225070 | 215447 | 214433 | 216206 | 215721 | 216849 | 218445 | 212134 | | 0 |
| 30 | Norway/Norvège | | 49751 | 49854 | 53947 | 53549 | 54803 | 53520 | 54241 | 54892 | 54153 | | 9 |
| 31 | Poland/Pologne | | 485407 | 453170 | 418883 | 405078 | 402108 | 387240 | 401569 | 396651 | 398952 | | -18 |
| 32 | Portugal | | 59921 | 71127 | 84586 | 82260 | 83469 | 88089 | 82952 | 84659 | 85538 | | 43 |
| 33 | Slovak Rep./Rép. slovaque | | 72051 | 52548 | 50368 | 47448 | 50645 | 48741 | 49082 | 48595 | 47866 | | -34 |
| 34 | Spain/Espagne | | 287366 | 318370 | 370243 | 384419 | 384811 | 402171 | 409488 | 425236 | 440649 | | 53 |
| 35 | Sweden/Suède | | 72191 | 73747 | 69832 | 68315 | 68976 | 69955 | 70726 | 69688 | 66955 | | -7 |
| 36 | Switzerland/Suisse | | 52749 | 51044 | 52488 | 51709 | 52548 | 51582 | 52578 | 53036 | 53636 | | 2 |
| 37 | Turkey/Turquie | * | 170059 | 220720 | 256776 | 279957 | 262099 | 270617 | 286283 | 296602 | 312312 | | 84 |
| 38 | UK/Royaume-Uni | | 771415 | 710129 | 672091 | 673967 | 677020 | 656921 | 662688 | 660424 | 657396 | | -15 |
| 39 | North America/Amérique N | * | 6825000 | 7206600 | 7609400 | 7846800 | 7728800 | 7767600 | 7834200 | 7937100 | 7988400 | | 17 |
| 40 | OECD/OCDE Europe | * | 5385400 | 5213900 | 5181200 | 5203900 | 5236300 | 5193400 | 5300000 | 5309300 | 5290800 | | -2 |
| 41 | EU15/UE15 | * | 4257700 | 4149700 | 4123900 | 4136100 | 4182000 | 4157000 | 4224900 | 4229500 | 4194300 | | -1 |
| 42 | OECD/OCDE | * | 13962600 | 14273200 | 14674500 | 14966200 | 14869700 | 14900800 | 15084700 | 15202100 | 15241600 | | 9 |
| 43 | | | | | | | | | | | | | |
| 44 | Notes: | | | | | Notes : | | | | | | | |
| 45 | a) Total aggregate anthropogenic emissions of $CO_2$, $CH_4$, $N_2O$, HFCs, PFCs and $SF_6$, excluding emissions/removals from land-use change and forestry. | | | | | a) Total anthropogène agrégé des émissions de $CO_2$, $CH_4$, $N_2O$, HFC, PFC et $SF_6$, excluant les émissions/puits issues du changement de l'utilisation des sols et de la sylviculture. | | | | | | | |
| 46 | | | | | | | | | | | | | |
| 47 | | | | | | | | | | | | | |
| 48 | | | | | | | | | | | | | |
| 49 | MEX) % of change: 2002/1990. | | | | | MEX) % variation : 2002/1990. | | | | | | | |
| 50 | KOR) National data; $CO_2$ and $CH_4$ only; % of change: 2004/1990. | | | | | KOR) Données nationales ; $CO_2$ et $CH_4$ uniquement; % variation : 2004/1990. | | | | | | | |
| 51 | | | | | | | | | | | | | |
| 52 | DNK) Excludes Faroe Islands and Greenland. | | | | | DNK) Non compris le Groenland et les îles Féroé. | | | | | | | |
| 53 | FRA) Metropolitan and overseas France. | | | | | FRA) Métropole et Outre-mer. | | | | | | | |
| 54 | TUR) National data. | | | | | TUR) Données nationales. | | | | | | | |
| 55 | TOT) Rounded figures, excludes Mexico and Korea. | | | | | TOT) Chiffres arrondis, exclut le Mexique et la Corée. | | | | | | | |

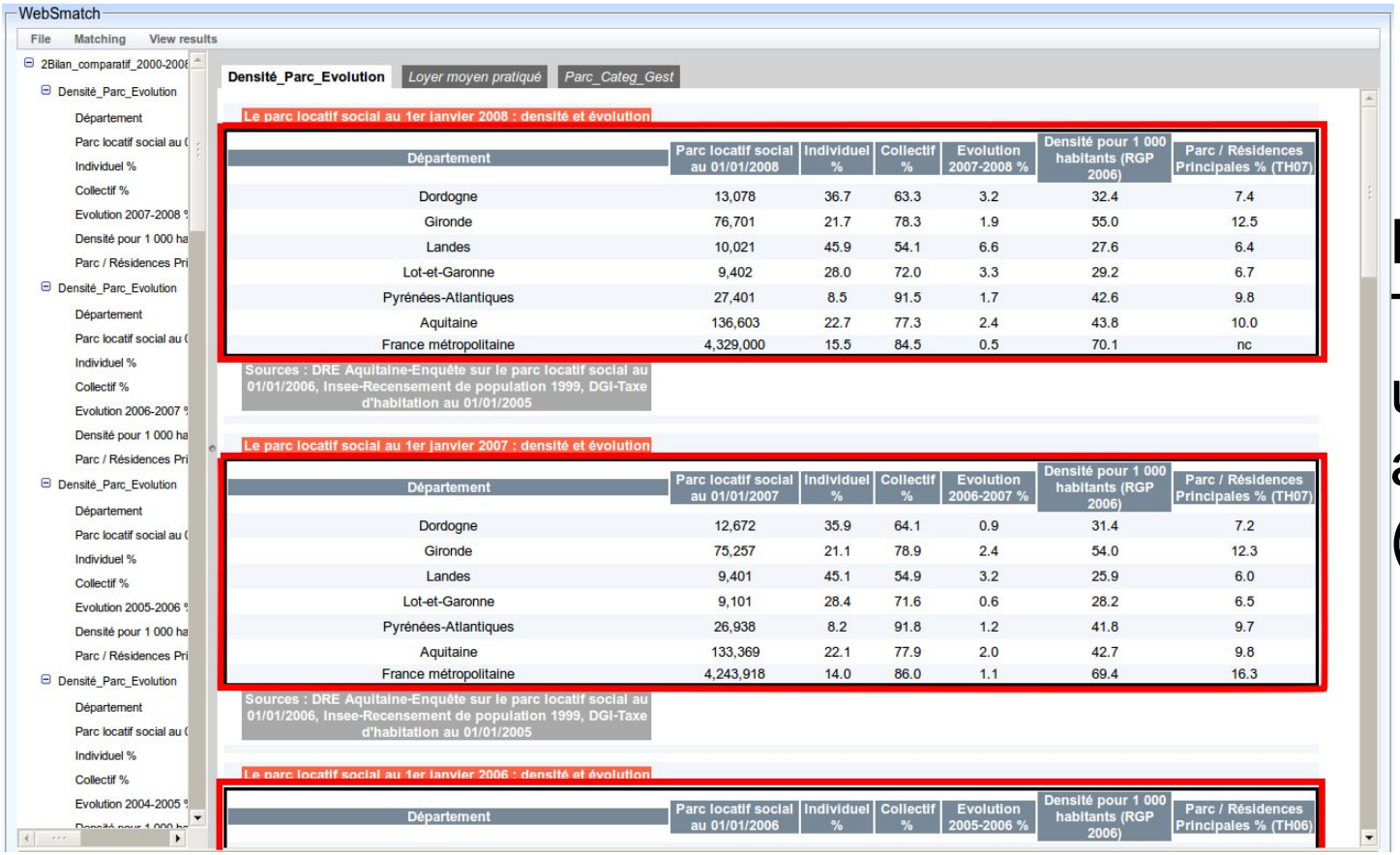
Find metadata such as titles

Identify collections for bidimensionnal tables

# WebSmatch metadata extraction service

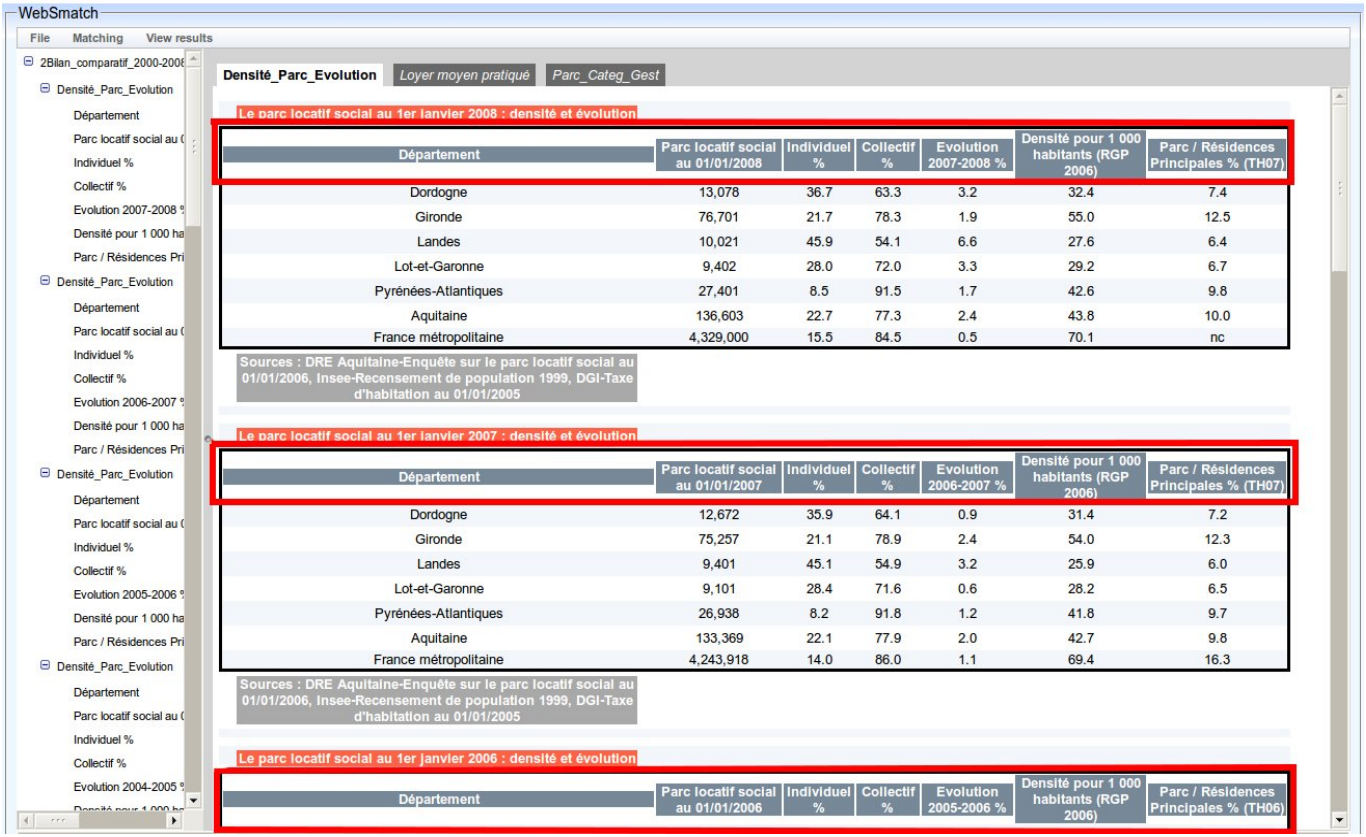


# MetaData Extraction: XLS example



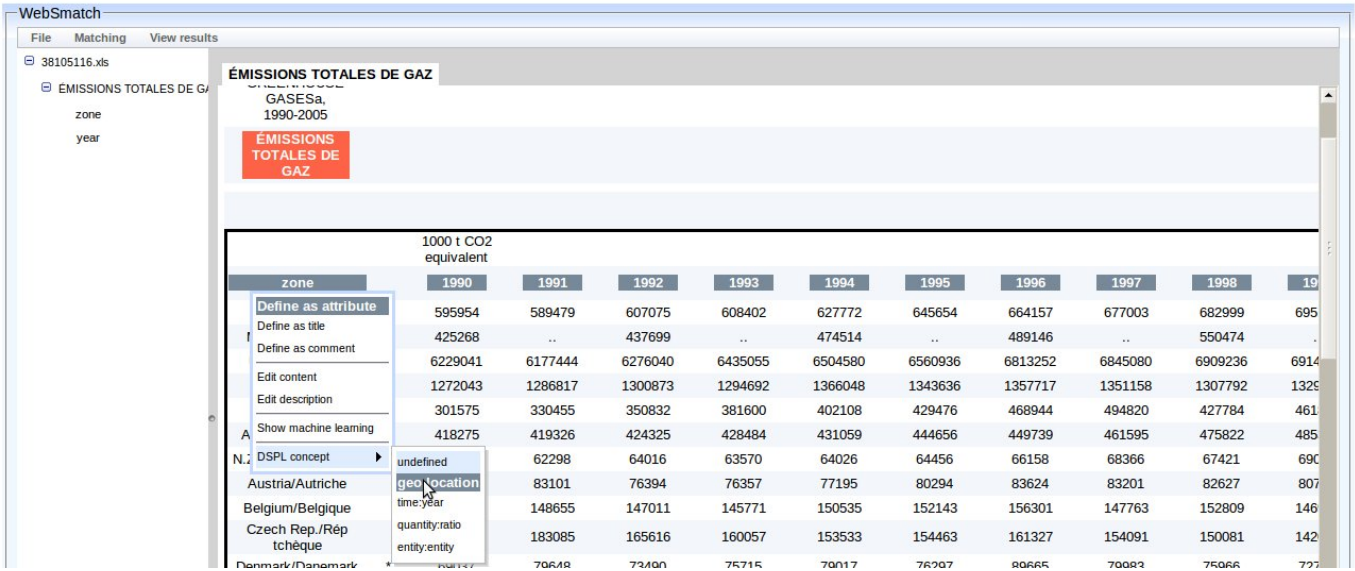

**First step :**
Table detection using vision algorithms (dilate/erode)

# MetaData Extraction: XLS example




**Second step :**
Attribute detection using machine learning on cell content and neigboorhood

# MetaData Extraction: XLS example

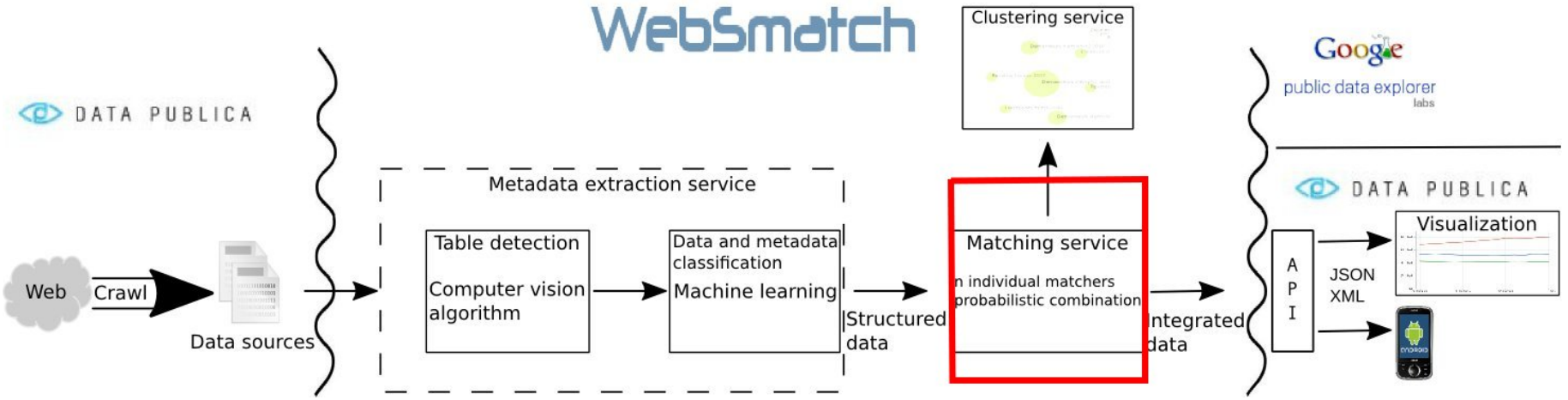



**Third step :** automatic detection of concepts using our YAM ++ ontology matching tool

*YAM++ came 1st at OAEI 2011 : http://oaei.ontologymatching.org/2011/results/*

# WebSmatch matching service

Relies on YAM++, which combines different metrics (String, Wordnet, Instance based) and matching techniques



# Data Visualization

Structured export formats easy to use for third parties

We use **DSPL (**DataSet Publishing Language de Google)

For bidimensionnal tables, we need to denormalize the data as DSPL uses flat CSV files

| | 1990 | 1995 | 1999 | 2000 | 2001 | 2002 |
|---|---|---|---|---|---|---|
| Germany/Allemagne | 1227860 | 1095654 | 1020669 | 1019764 | 1036736 | 1017514 |
| Spain/Espagne | 287366 | 318370 | 370243 | 384419 | 384811 | 402171 |
| Iceland/Islande | 3352 | 3138 | 3739 | 3684 | 3671 | 3684 |
| Ireland/Irlande | 55374 | 59372 | 67317 | 69127 | 70923 | 68971 |
| Italy/Italie | 516851 | 530264 | 546311 | 551594 | 557598 | 557816 |
| Luxembourg | 12687 | 9775 | 9002 | 9548 | 9830 | 10778 |
| Netherlands/Pays-Bas | 212963 | 225070 | 215447 | 214433 | 216206 | 215721 |
| Norway/Norvège | 49751 | 49854 | 53947 | 53549 | 54803 | 53520 |
| Poland/Pologne | 485407 | 453170 | 418883 | 405078 | 402108 | 387240 |
| Portugal | 59921 | 71127 | 84586 | 82260 | 83469 | 88089 |
| Slovak Rep./Rép. slovaque | 72051 | 52548 | 50368 | 47448 | 50645 | 48741 |
| Spain/Espagne | 287366 | 318370 | 370243 | 384419 | 384811 | 402171 |

⇒

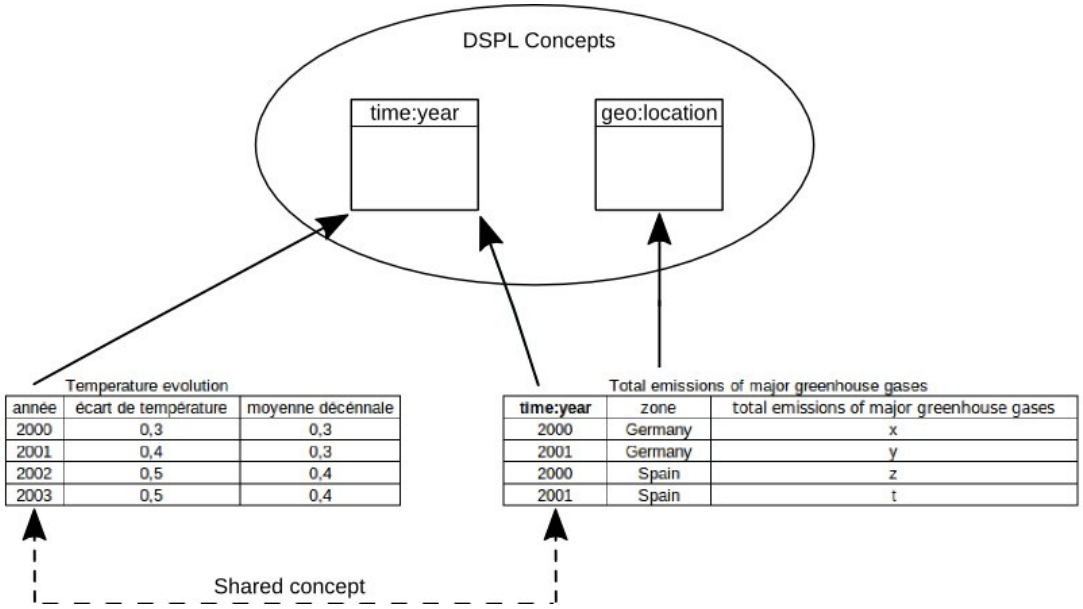
Total emissions of major greenhouse gases

| time:year | zone | total emissions of major greenhouse gases |
|---|---|---|
| 2000 | Germany | x |
| 2001 | Germany | y |
| 2000 | Spain | z |
| 2001 | Spain | t |

# Exporting the Results : integrated metadata

Combining datasets : aggregation or intersection

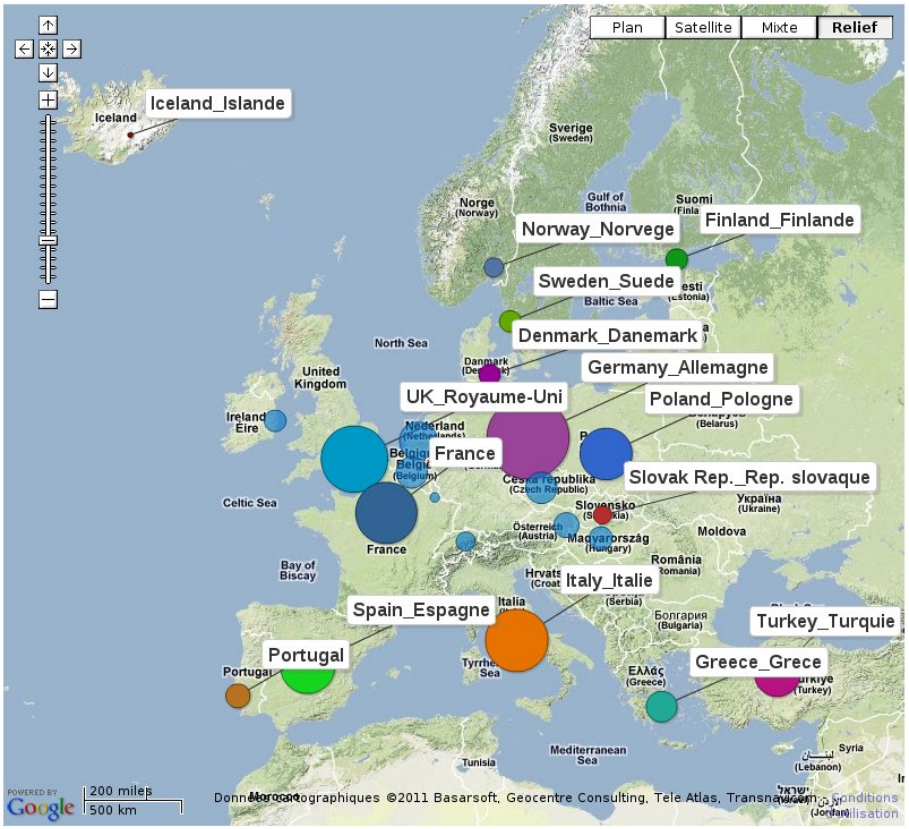
- using generic concepts such as time or location
- find a specific concept using the matching

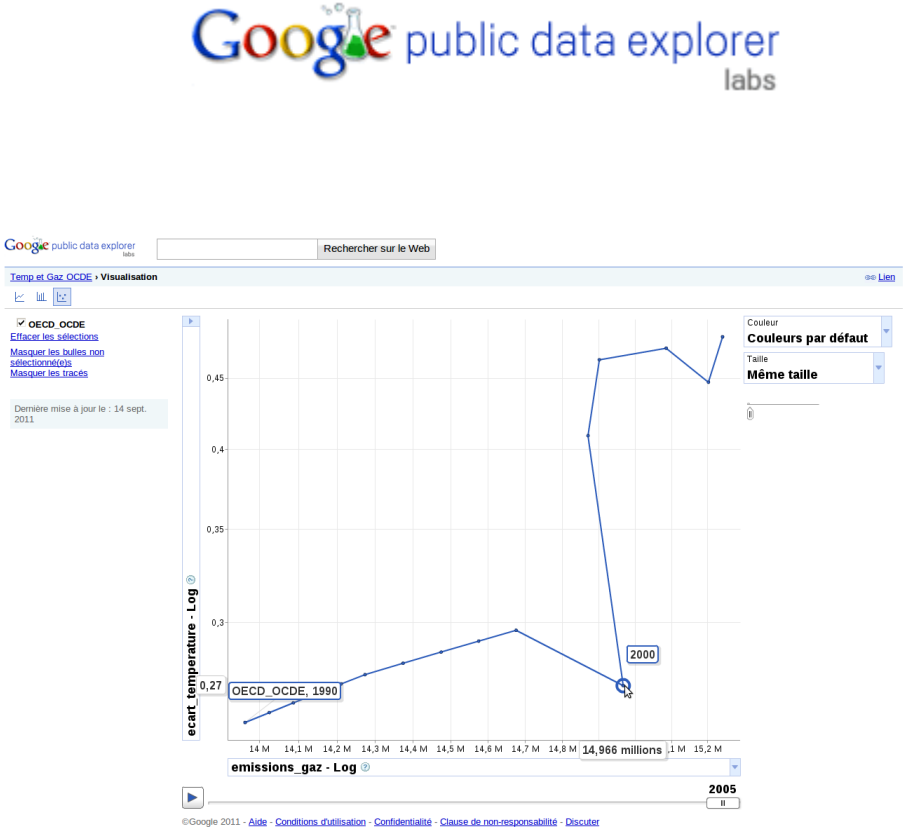



| année | écart de température | moyenne décénnale |
|---|---|---|
| 2000 | 0,3 | 0,3 |
| 2001 | 0,4 | 0,3 |
| 2002 | 0,5 | 0,4 |
| 2003 | 0,5 | 0,4 |

| time:year | zone | total emissions of major greenhouse gases |
|---|---|---|
| 2000 | Germany | x |
| 2001 | Germany | y |
| 2000 | Spain | z |
| 2001 | Spain | t |

# Visualizing the Results



# Visualizing the Results

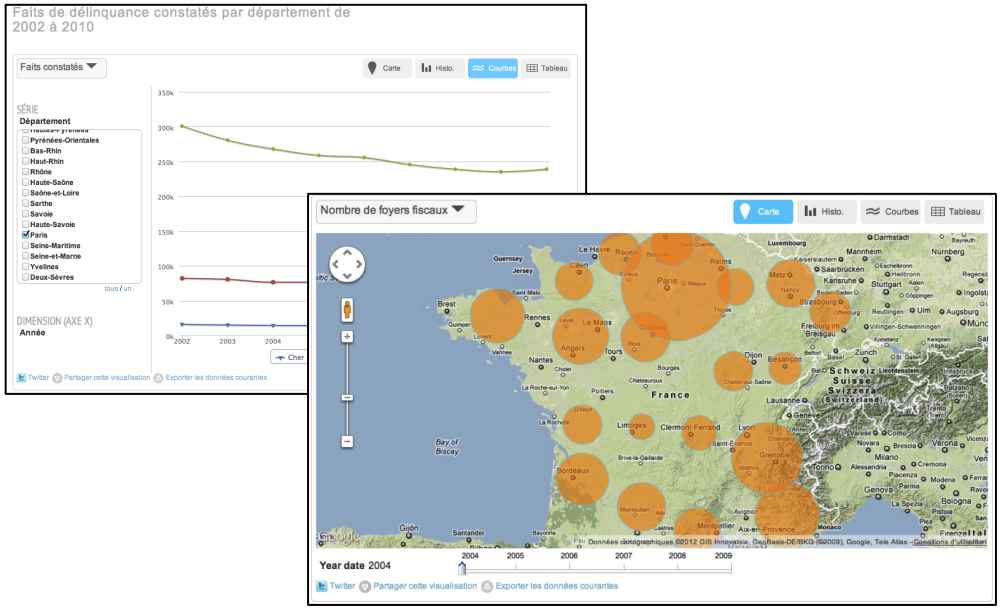
DATA PUBLICA

```
http://api.data-publica.com/…/content.json?
limit=10&filter={revenue_fiscal_par_foyer:{$gt:25000}}
```

Show Original


```
- [
  - {
      "region": "Ile de france",
      "rp_montant": 36063493357,
      "ts_nombre_foyers": 4916338,
      "ts_montant": 143628000000,
      "revenu_fiscal": 190090000000,
      "nombre_foyers_imposable": 4254456,
      "impot": 17776583634,
      "revenu_fiscal_foyers_imposables": 168618000000,
      "taux_foyers_imposable": 63.4,
      "nombre_foyers": 6714986,
      "rp_nombre_foyers": 1639891,
      "year": "Mon Jan 01 00:00:00 CET 2007",
      "revenue_fiscal_par_foyer": 28308.29964
    },
  - {
      "region": "Ile de france",
      "rp_montant": 37756320449,
      "ts_nombre_foyers": 4950056,
      "ts_montant": 149317000000,
      "revenu_fiscal": 199710000000,
      "nombre_foyers_imposable": 4344660,
      "impot": 19120373742,
      "revenu_fiscal_foyers_imposables": 178638000000,
      "taux_foyers_imposable": 64.2,
      "nombre_foyers": 6766538,
      "rp_nombre_foyers": 1668476,
      "year": "Tue Jan 01 00:00:00 CET 2008",
      "revenue_fiscal_par_foyer": 29514.38595
```


- Multi format (json, xml, spreadsheet, csv)
- Geolocalized queries
- Mashups



# Perspectives

1. Scaling up to high numbers of data sources (tens of thousand) through incremental extraction
2. Clustering documents on specific concepts & concept instances
3. Integration with other tools
    - Google Refine
    - RDF export

# References

Remi Coletta, Emmanuel Castanier, Patrick Valduriez, Christian Frisch, DuyHoa Ngo, Zohra Bellahsene. WebSmatch : a platform for data and metadata integration. First Int. Workshop on Open Data, Nantes, May 2012.

http://websmatch.gforge.inria.fr/

http://www.youtube.com/watch?v=sqeU1lkXW_A