Simuler les données manquantes dans les Open Data ?

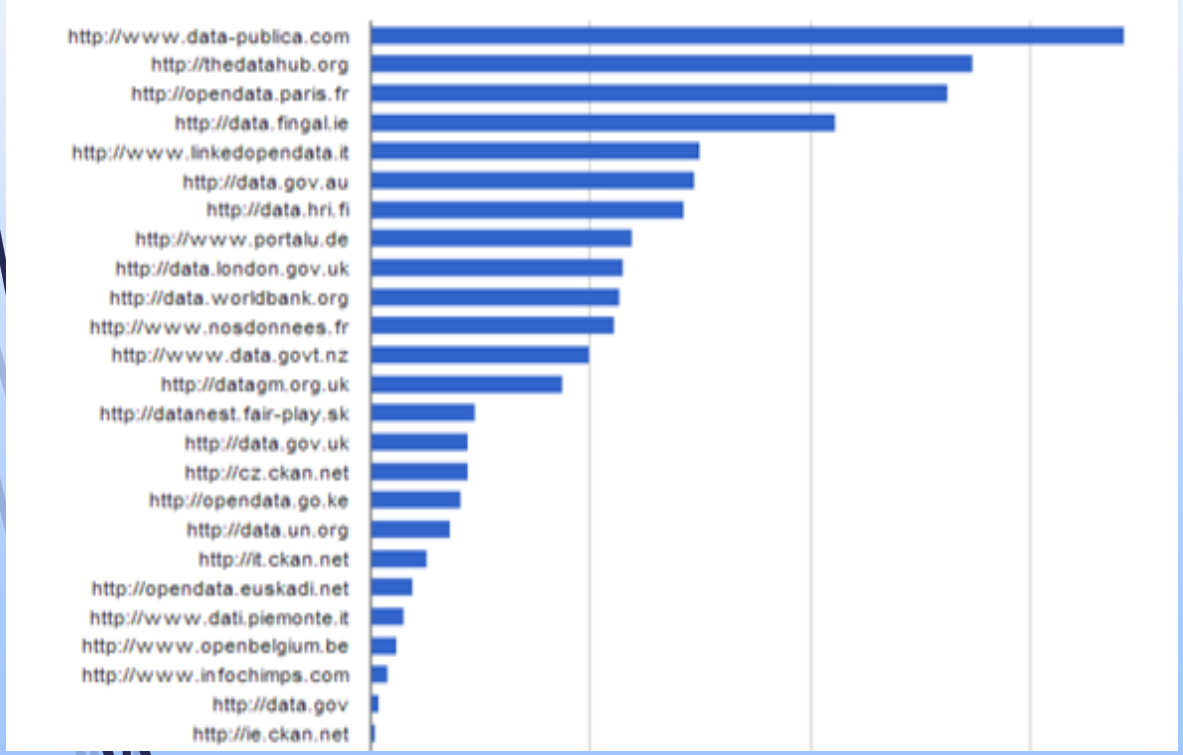
Présenté par : Imen Megdiche
Directeur de Thèse : Olivier Teste
Co-directeur de thèse : Alain Berro

Plan

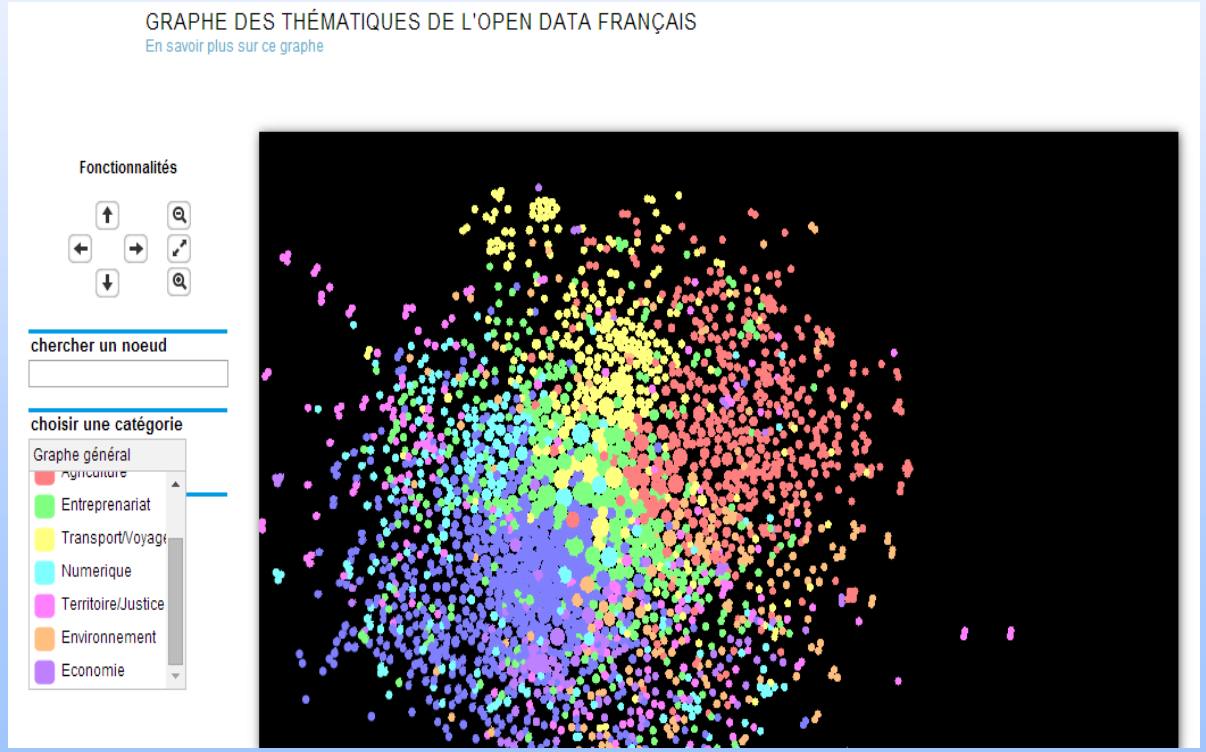
1. Impact des Open Data sur les SI
2. Processus d'entreposage d'Open Data
3. Données manquantes dans les Open Data
4. Simuler les données .. une bonne alternative ?

Impact des Open Data sur les SI

Accroissement



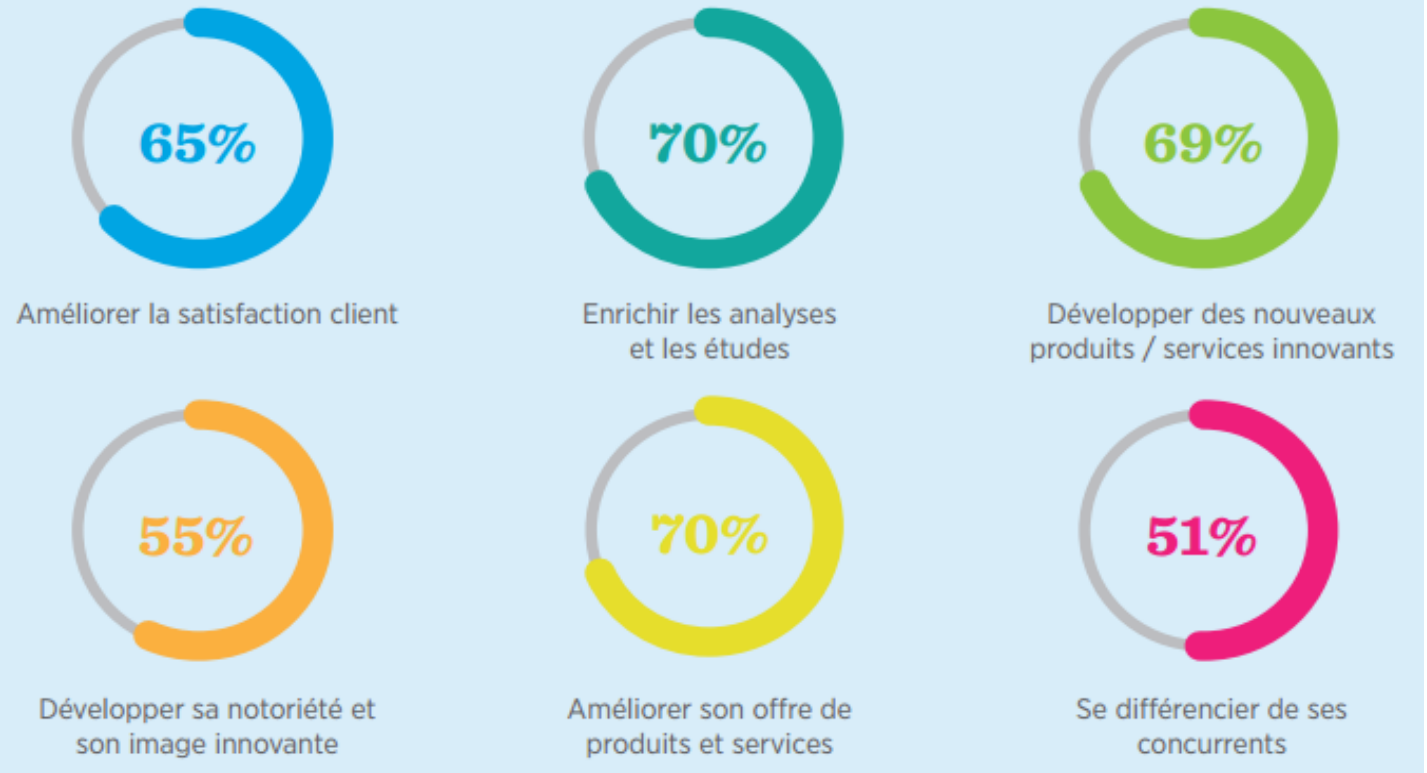
Diversité & Richesse



http://wwwdb.inf.tu-dresden.de/opendatasurvey/

Impact des Open Data sur les SI

Quelques témoignages sur les principaux bénéfices attendus de la réutilisation des Open Data



Impact des Open Data sur les SI

MAIS ….

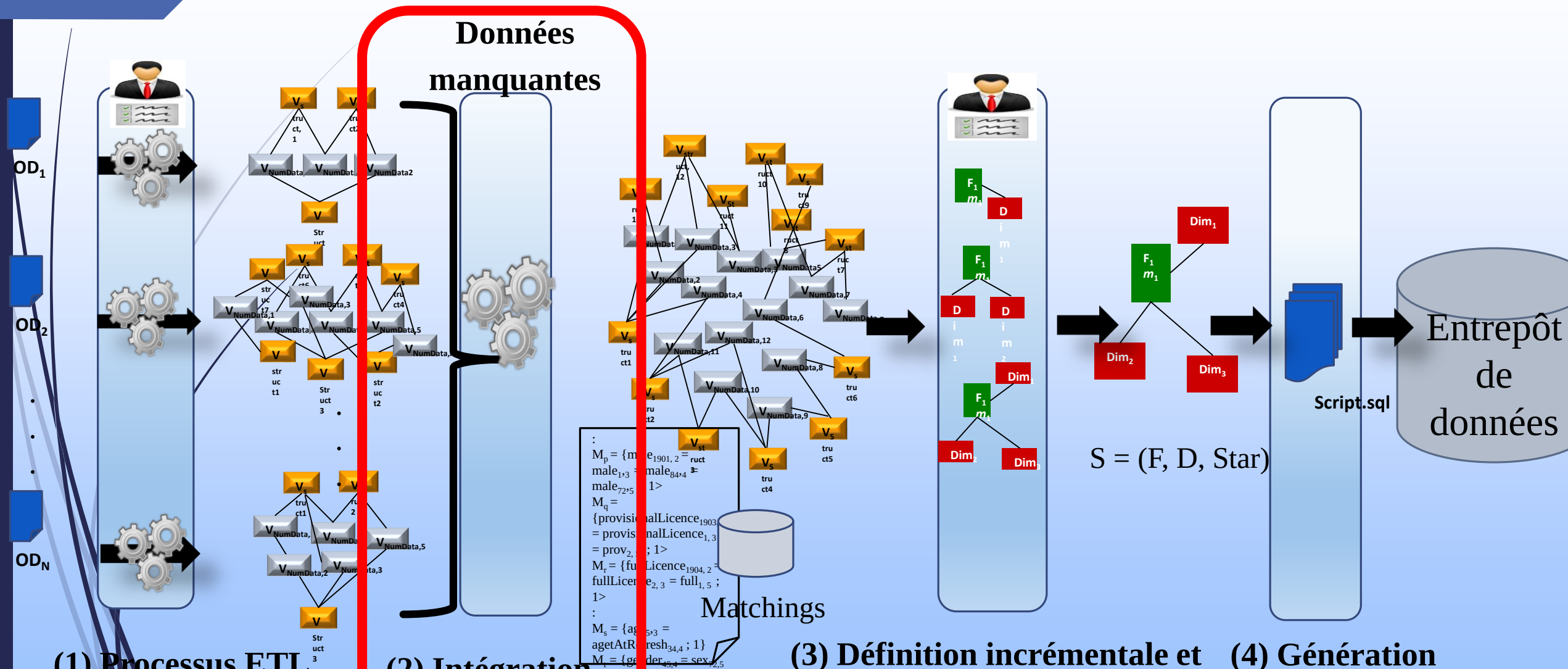
*** Il faut les chercher ⬅ dispersés sur plusieurs fournisseurs**

*** Il faut les nettoyer et aligner ⬅ hétérogénéité sémantique et structurelle**

*** Il faut les compléter et corriger ⬅ données manquantes et erronées**

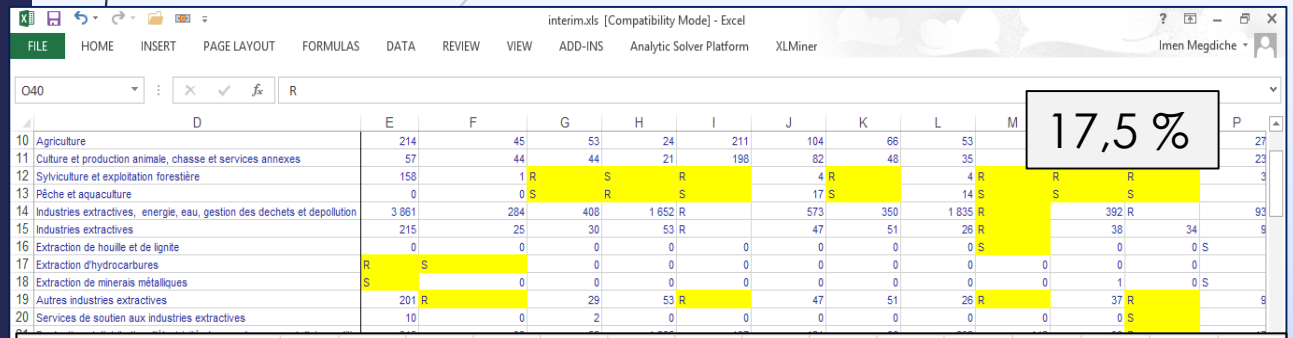
Intégrer les Open Data dans les entrepôts de données ?

Processus d'entreposage des Open Data



Données manquantes dans les Open Data

…issues de la qualité



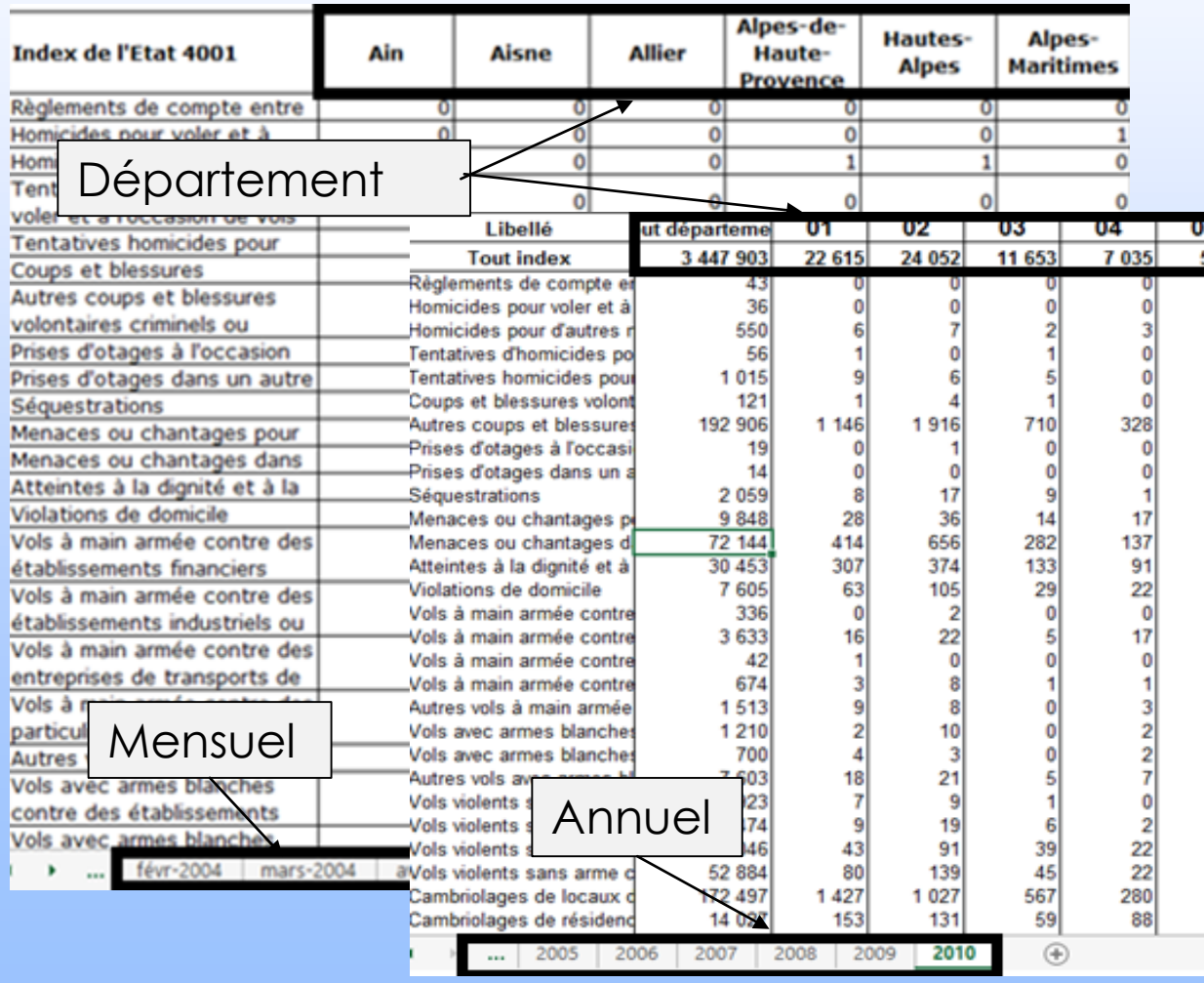
	D	E	F	G	H	I	J	K	L	M	N	O	P
10	Agriculture	214	45	53	24	211	104	66	53				27
11	Culture et production animale, chasse et services annexes	57	44	44	21	198	82	48	35				23
12	Sylviculture et exploitation forestière	158	1	R S	R		4	R	4	R	R	R	3
13	Pêche et aquaculture	0	0	S R	S		17	S	14	S	S	S	
14	Industries extractives, energie, eau, gestion des dechets et depollution	3 861	284	408	1 652	R	573	350	1 835	R	392	R	93
15	Industries extractives	215	25	30	53	R	47	51	26	R	38	34	9
16	Extraction de houille et de lignite	0	0	0	0	0	0	0	0	S	0	0 S	
17	Extraction d'hydrocarbures	R S		0	0	0	0	0	0	0	0	0	
18	Extraction de minerais métalliques	S 0		0	0	0	0	0	0	0	1	0 S	
19	Autres industries extractives	201	R	29	53	R	47	51	26	R	37	R	9
20	Services de soutien aux industries extractives	10	0	2	0	0	0	0	0	0	0	S	

	1990	1995	1996	1997	1998	1999	2000	2001*	2002	2003	2004 (r)	2005 (r)	2006 (r)	2007 (r)	2008
BTS et assimilés	53 113	77 929	77 255	80 574	85 204	88 421	95 530	99 273	103 629	103 497	108 839	102 544	102 154	101 389	105 931
DUT	27 835	37 380	39 466	39 804	44 256	46 701	47 478	47 997	48 877	48 142	47 018				46 714
DEUG, DEUST	88 353	132 390	134 601	135 498	137 381	131 756	124 700	121 833	120 935	119 017	110 171				1 748
DEUG intermédiaires	///	///	///	///	///	///	///	///	///	///	///				///
Licence	71 325	127 178	131 139	135 490	138 102	135 306	135 017	129 191	130 188	133 437	123 314				824
Licence LMD	///	///	///	///	///	///	///	///	///	///	13 993	82 700	128 168	127 232	123 465
Licence professionnelle	///	///	///	///	///	///	///	3 620	8 013	12 921	17 159	23 909	30 090	34 915	37 665
Maîtrise	51 169	80 833	86 736	88 333	92 086	93 734	93 304	95 752	96 034	97 178	94 146	44 088	9 184	4 433	1 886
Maîtrises intermédiaires	///	///	///	///	///	///	///	///	///	///	///	54 137	88 674	89 200	///
Master professionnel et DESS	12 624	21 077	23 463	24 866	26 853	28 885	32 612	38 094	43 256	47 174	48 939	57 644	63 995	65 699	65 221
Master recherche et DEA	20 024	25 420	26 073	24 481	23 910	23 520	23 428	24 503	26 529	26 819	27 233	26 535	24 703	23 228	22 140
Master indifférencié	///	///	///	///	///	///	///	///	///	///	///	647	2 985	5 020	7 069
Doctorat	7 161	8 969	9 448	10 253	9 597	9 467	9 991	9 011	8 243	8 087	8 931	9 277	10 045	10 664	10 678
Diplôme de santé (docteur)	8 797	7 717	7 721	7 942	6 824	6 946	6 661	nd	5 755	7 185	6 388	6 844	6 790	7 174	6 795
Capacité en médecine	nd	1 838	1 816	2 125	1 924	1 337	1 554	nd	1 900	1 882	2 072	2 185	2 190	1 924	1 882
DES, DIS, DESC (1)	nd	3 540	3 740	3 830	3 565	3 107	3 224	nd	2 785	2 741	2 366	2 489	2 236	2 974	3 700
Diplôme d'ingénieur	16 080	21 851	22 689	22 828	23 068	23 658	24 624	26 023	26 155	26 437	26 817	27 638	27 676	27 520	28 619
Diplôme des écoles de commerce	12 151	18 358	18 373	16 669	16 424	15 993	18 342	20 684	21 440	24 363	25 179	25 066	25 626	24 397	22 246

21,1 %

* : rupture de série : le mode de collecte de l'information sur les diplômes a évolué entre 2000 et 2001, en particulier pour les doctorats.
nd : donnée non disponible.
/// : absence de donnée due à la nature des choses.
r : données révisées.
(1) : DES, DESC : diplôme d'études spécialisées (complémentaires) ; DIS : Diplôme interdisciplinaire de spécialisation.
Champ : France.
Source : Depp.

….issues de l'intégration/alignement



Simuler les données … une bonne alternative ?

- \+ Donner une vision complète des données
- \+ Construire et croiser le plus de sources possibles pour avoir des scénarios d'analyses riches
- \+ Gagner le temps
- − Biaiser les données
- − Informer les utilisateurs que les données sont simulées