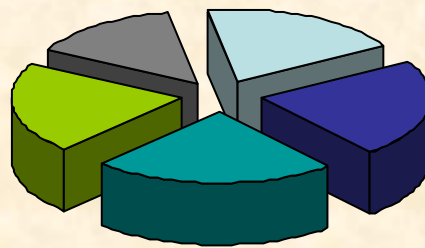


Wikipedia content

thematic analysis

1. Subject and scope of research

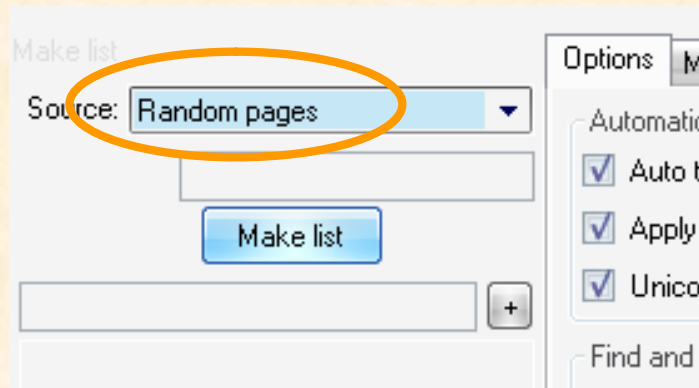
- **Jimbo Wales:**
- “Imagine if every person on earth could share in the free and total access to all human knowledge ...”
- **Question:** What is the structure of human knowledge ?
- We will try to analyze it on Wikipedia



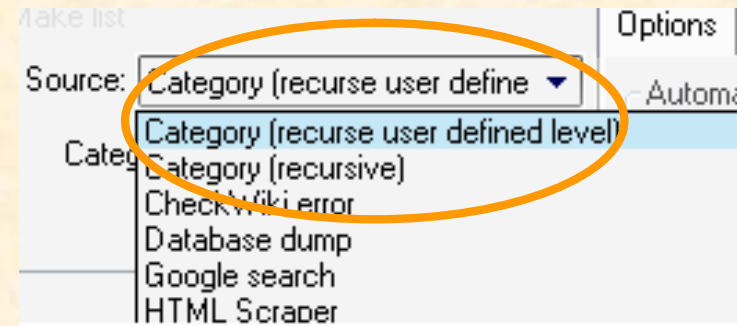
fields of knowledge

2. Metodology

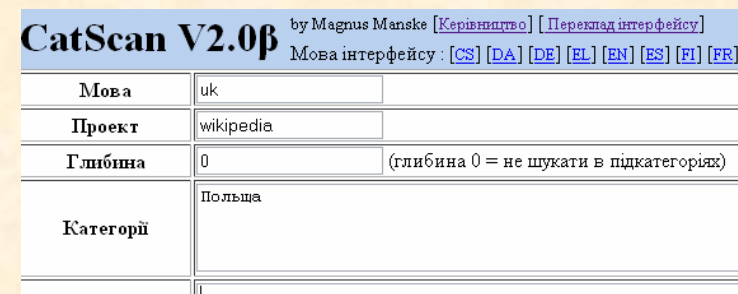
- Random selection
- Par example Using AWB



- Recursive analysis of categories Using AWB



- Or ScanCat



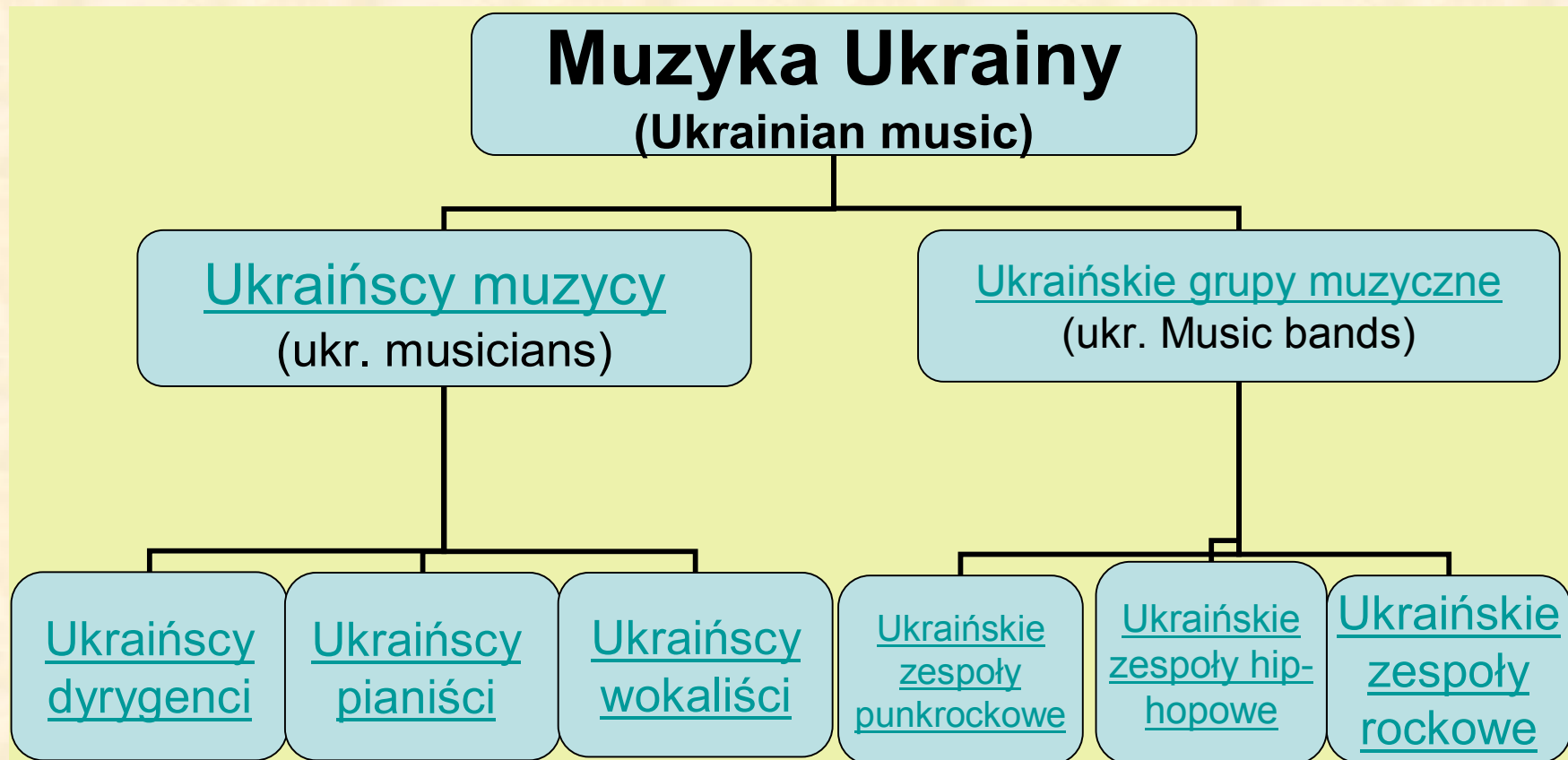
2.1 Random selection

sample and error	En 3,5M+	Fr, De 1M+	Ru, PL 800K+	Uk 300K+	Cs 200K+
100 art. 1%	35K	10K	8K	3K	2K
250 art. 0,4%	14K	4K	3,2K	1,2K	800
1000 art. 0,1%	3,5K	1K	800	300	200

2.2 Recursive category analysis –

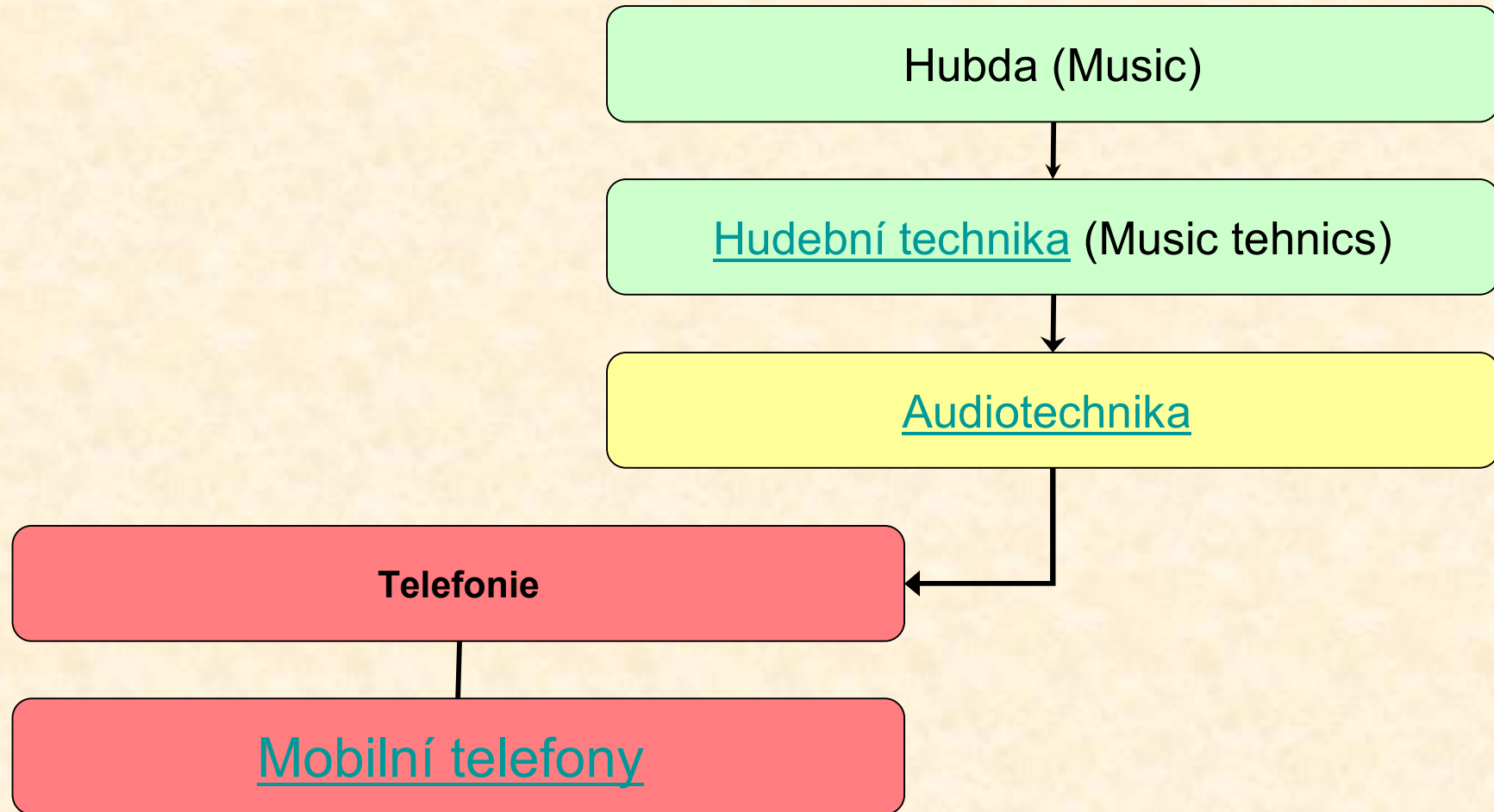
Categorization transitive, subsets according to the content

Polish Wikipedia, Kategoria:Muzyka Ukrainy



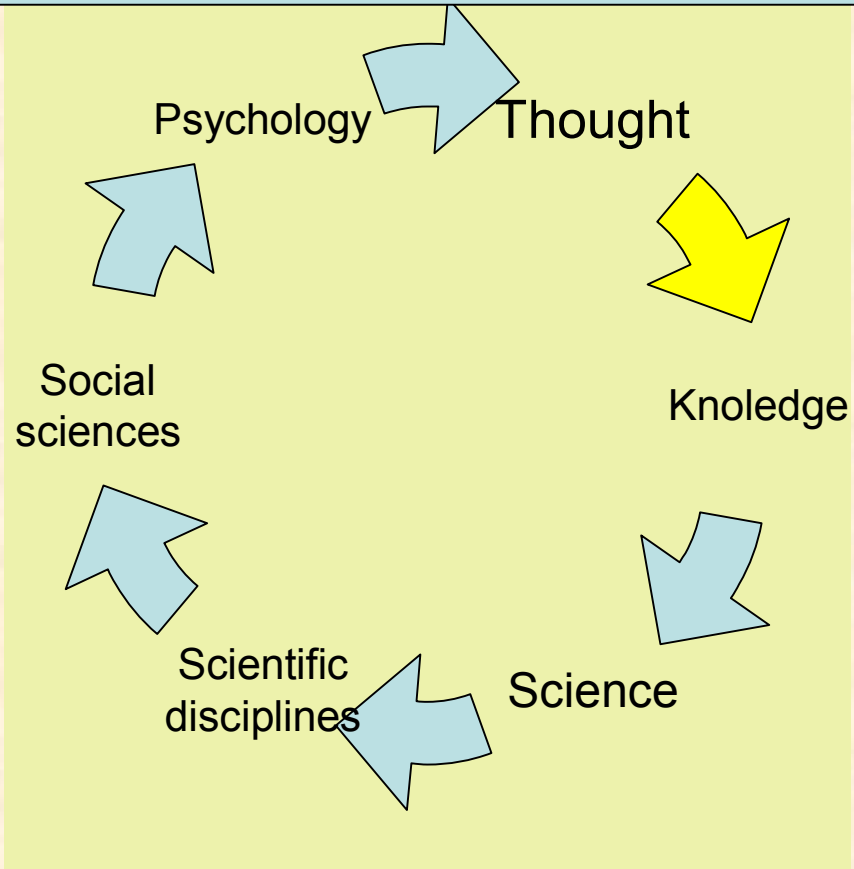
2.3 Recursive analysis –

Categorization transitive, subsets not according to the content
Czech Wikipedia, *Kategorie:Hudba*

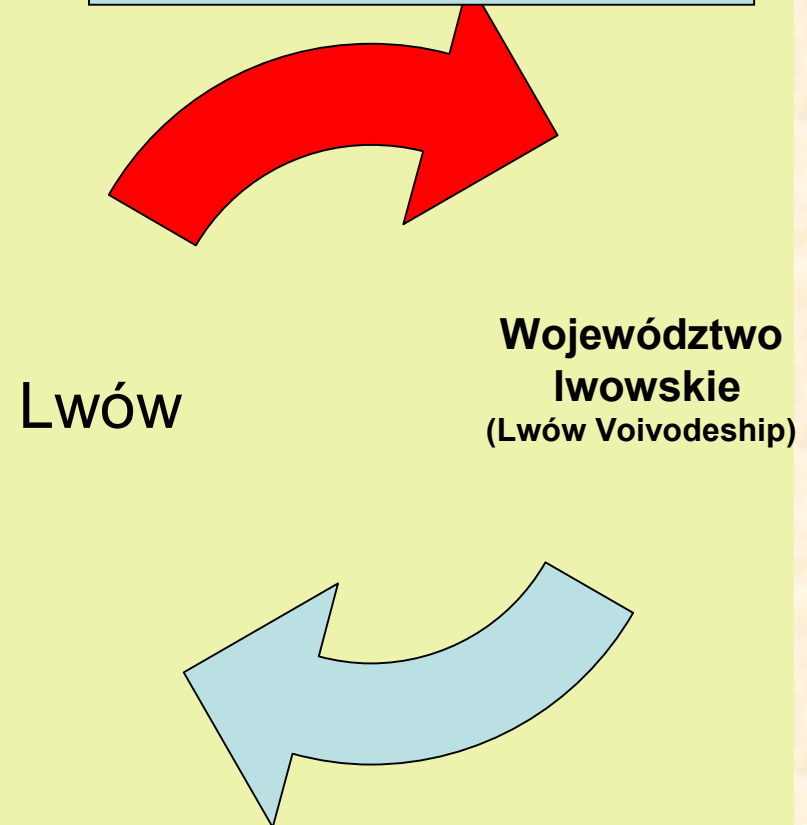


2.4 Recursive analysis – Not transitive categorization

Classical untransitivity (En wiki)



Cycle in Pl-wiki (now fixed...)

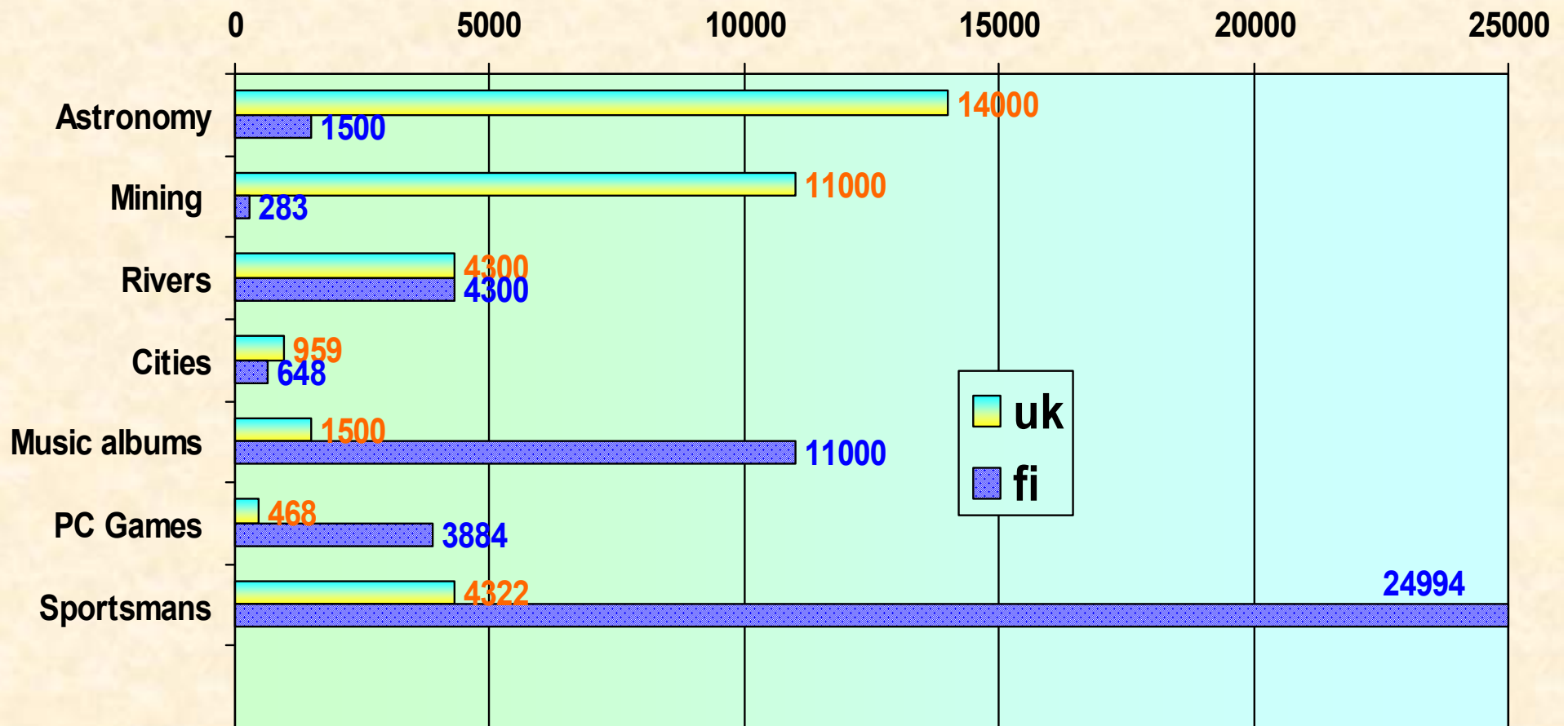


2.5 Resume

- Random selection – for broader areas
Recursive analysis – for narrower areas

3.1. special research – Ukr vs Fin

Comparison of Ukrainian and Finnish WP while they was equal
(~ 255K art.)



3.2. special research – Music

- Structure of music articles:

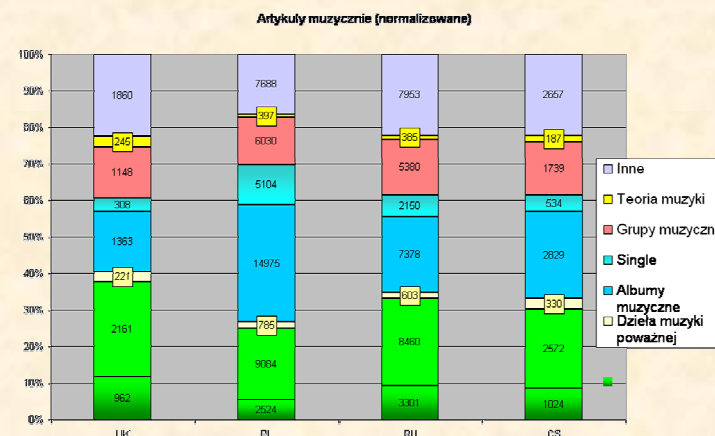
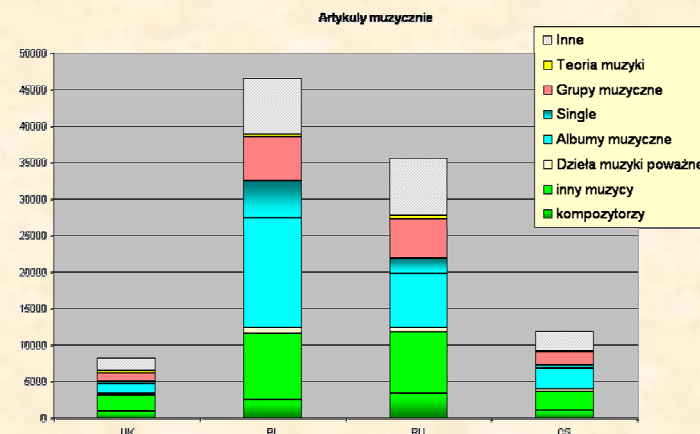
- Absolute bar chart
- Normilized bar chart
 - Green - musicians
 - Blue – albums & single
 - Blue – bands
 - Yellow – musicology

❖ UK (Українська)

❖ PL (Polska)

❖ RU (Русский)

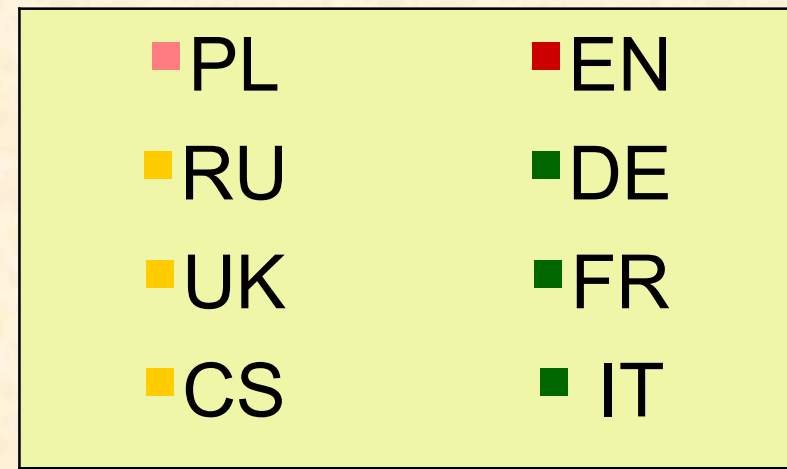
❖ CS (Česky)



4.1. General researches

1. Make a random sample
2. Analyze:
 - by subject
 - by type
3. Make diagrams

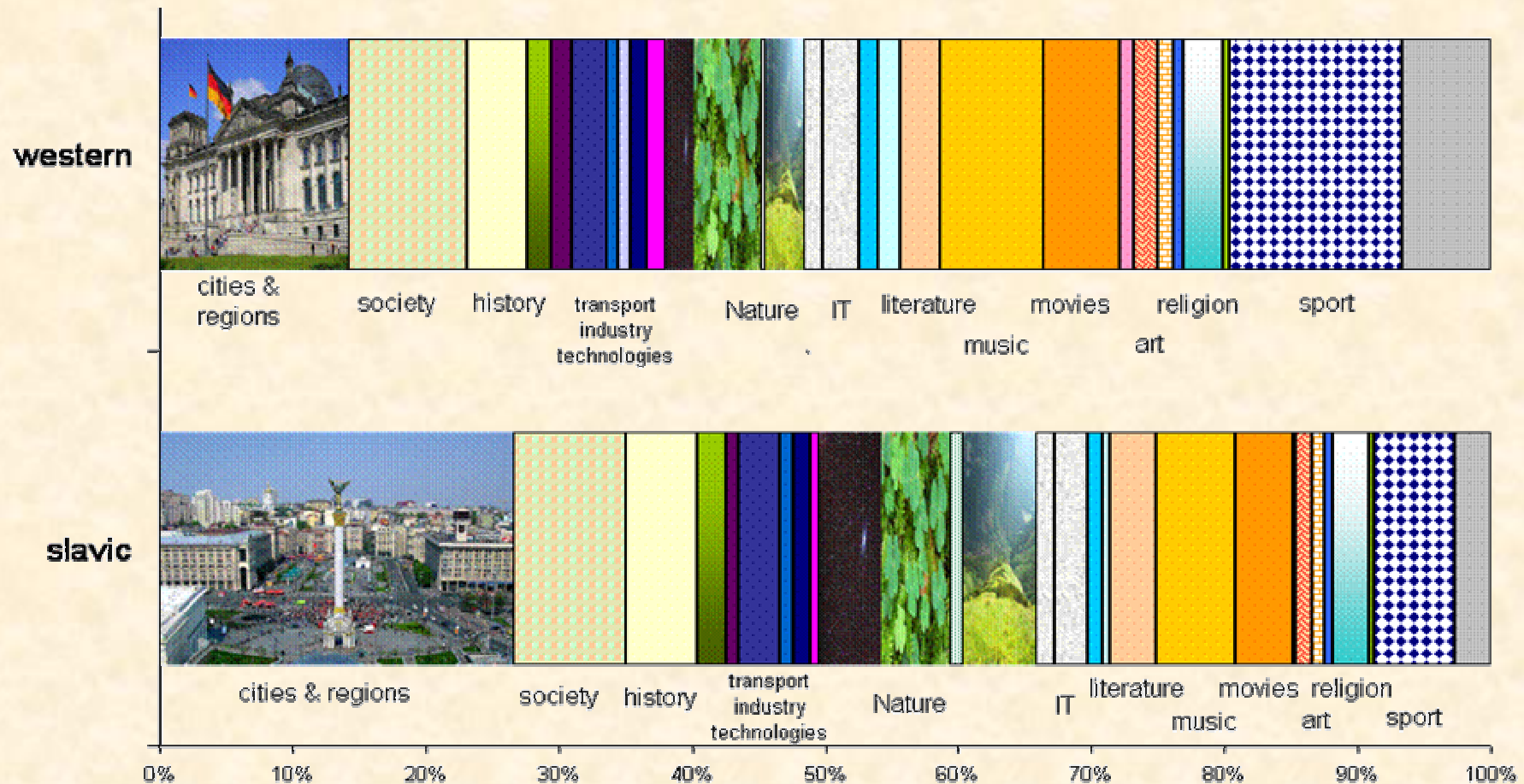
1	стаття	тематика	тип предмета	повторюв	повторюв	зведена таблиця			зведена таблиця		
2	Palác Michny z Vácinoва	архітектура	об'єкт	22	3	тематика	знайдено	частка	тематика	знайдено	частка
3	Vitraz	архітектура	поняття	51	3	соц. географія	27	населення	14.4%	поняття	51
4	Alfred Mosher Butts	архітектура	людина	41	3	суспільство	10	суспільство	5.3%	об'єкт	22
5	Sojuz TMA-17	астрономія	модель	8	4	історія	14	історія	7.4%	людина	41
6	Messier 67	астрономія	об'єкт	22	4	військо	2	військо	1.1%	організація	9
7	Německé středisko pro letectví	астрономія	організація	9	4	промисловість	2	виробництво	1.1%	населений пункт	19
8	Kap (měsíc)	астрономія	об'єкт	22	4	транспорт і зв'язь	6	транспорт	3.2%	адміністрація	7
9	Cardale Babington	біологія	людина	41	13	геологія та гірництво	2	геологія	1.1%	людина	8
10	Centrální dogma molekulární biologie	біологія	поняття	51	13	археологія	0	археологія	0.0%	персонаж	2
11	Kosatice bahenní	біологія	поняття	51	13	фізика	3	фізика	1.6%	дисамбіг	9
12	Dynaktin	біологія	поняття	51	13	хімія	1	хімія	0.5%	модель	8
13	Homo antecessor	біологія	поняття	51	13	астрономія	4	астрономія	2.1%	список	6
14	Koryši	біологія	поняття	51	13	біологія	13	біологія	6.9%	огляд	1
15	Mitozom	біологія	поняття	51	13	медицина	3	медицина	1.6%	споруда	0
16	Památný strom	біологія	поняття	51	13	фіз. географія	14	фізична географія	7.4%	тевір	11
17	Teigmentum	біологія	поняття	51	13	математика	5	математика	2.7%	видання	1
18	Dvouděložné	біологія	поняття	51	13	IT	10	IT	5.3%	інше	5
19	Sběrač rosný	біологія	поняття	51	13	філологія	2	філологія	1.1%		
20	Lewisuchus	біологія	поняття	51	13	освіта	1	освіта	0.5%		
21	Korandův smrk	біологія	об'єкт	22	13	література	8	література	4.3%		
22	Canadian War Museum	військо	об'єкт	22	2	музика	11	музика	5.9%		
23	Zahar Pnlepin	військо	людина	41	2	кіно	8	кіно і ТВ	4.3%		
24	Hydrodynamická vertikální zóna	геологія та гірництво	поняття	51	2	театр	1	театр	0.5%		
25	Richard Lydekker	геологія та гірництво	людина	41	2	живопис	2	живопис	1.1%		
26	1586	дата	список	6	3	архітектура	3	архітектура	1.6%		
27	1985	дата	список	6	3	грашка	3	грашки	1.6%		
28	Ochotný třjad	дисамбіг	дисамбіг	9	9	реалія	7	реалія	3.7%		
29	Horčická	дисамбіг	дисамбіг	9	9	міфологія	0	міфологія	0.0%	усього (без дати)	200
30	Obláhání Brna	дисамбіг	дисамбіг	9	9	спорт	17	спорт	9.0%		
31	DeSoto County	дисамбіг	дисамбіг	9	9	інше	9	інше	4.8%		
32	Letka	дисамбіг	дисамбіг	9	9	усього (без дати)	186		100.00%		
33	П	дисамбіг	дисамбіг	9	9	дисамбіг	9				
34	Imogta	дисамбіг	дисамбіг	9	9	дисамбіг	9				



4.2 Western and Slavic WP by subject

Western – more sports

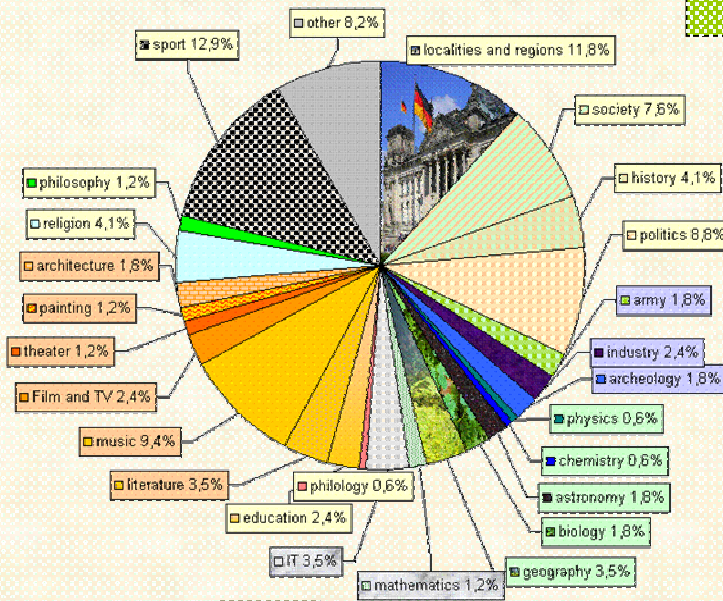
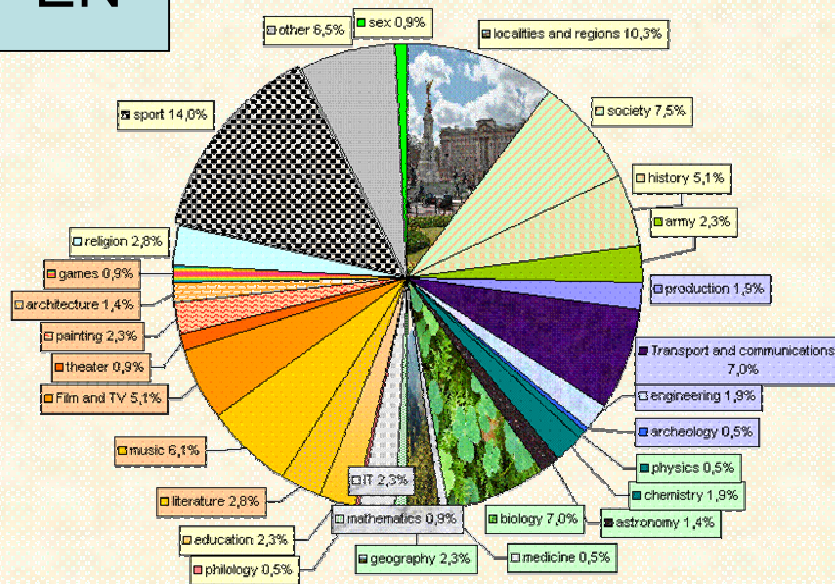
Slavic – more geography



4.3 Western Wikipedias by subject

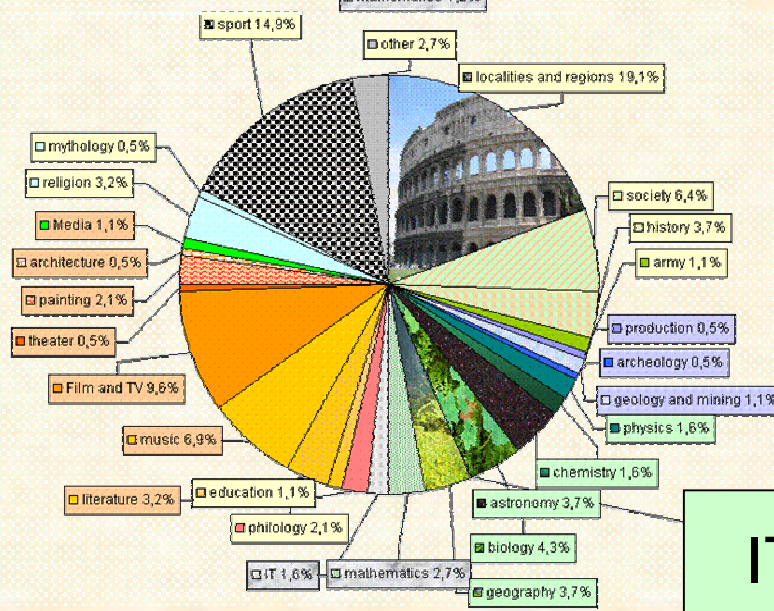
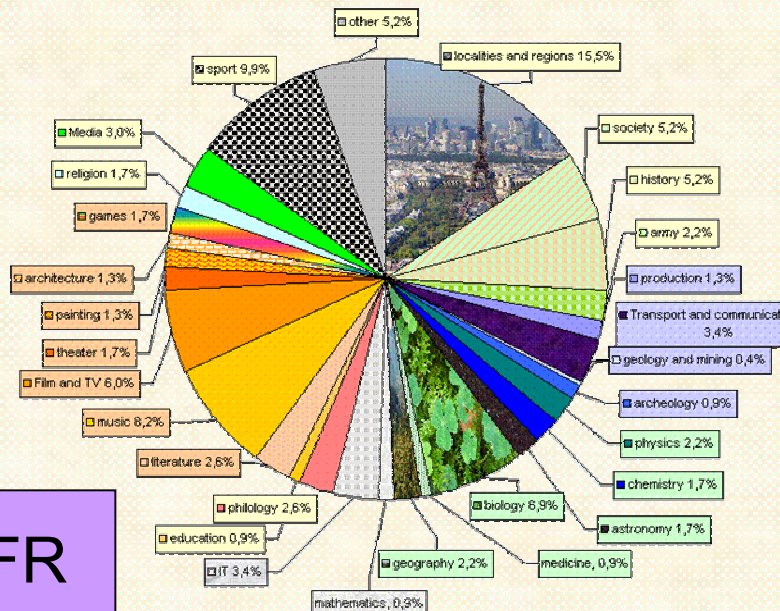
EN

DE



FR

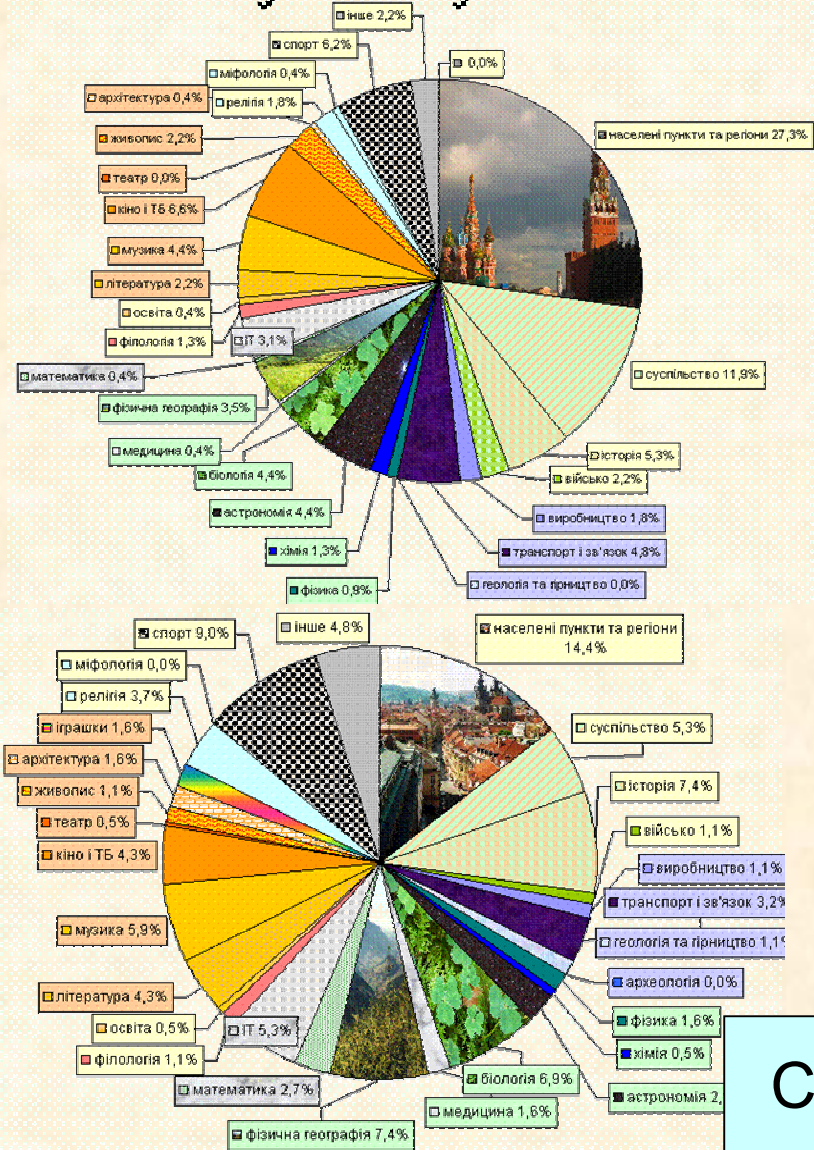
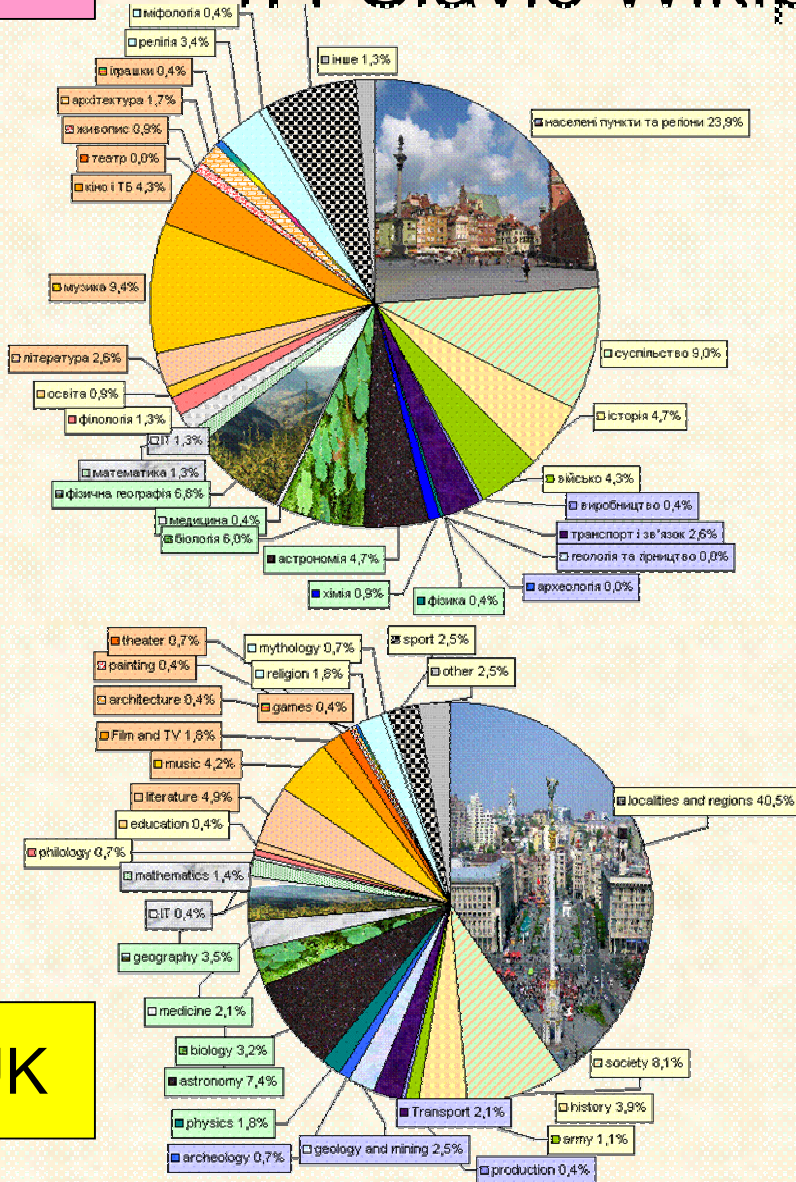
IT



PL

4.4 Slavic Wikipedias by subject

RU



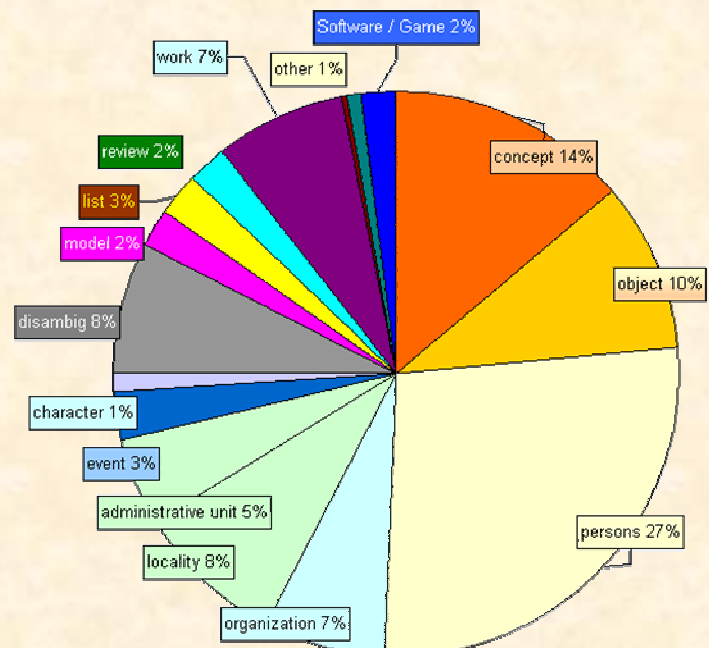
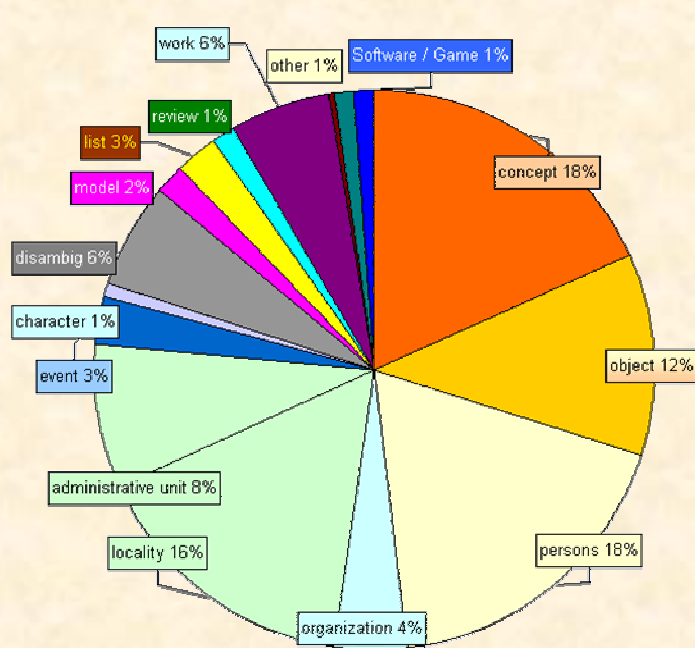
UK

CS

4.6

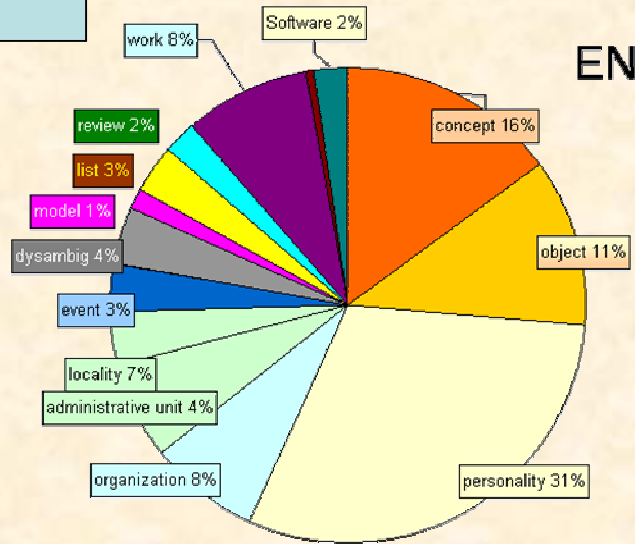
Slavic – more objects & concepts

Western – more persons & companies

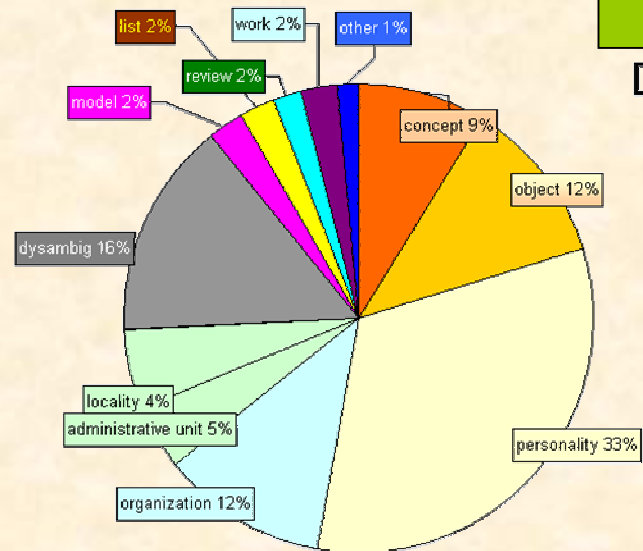


4.7 Western Wikipedias by type

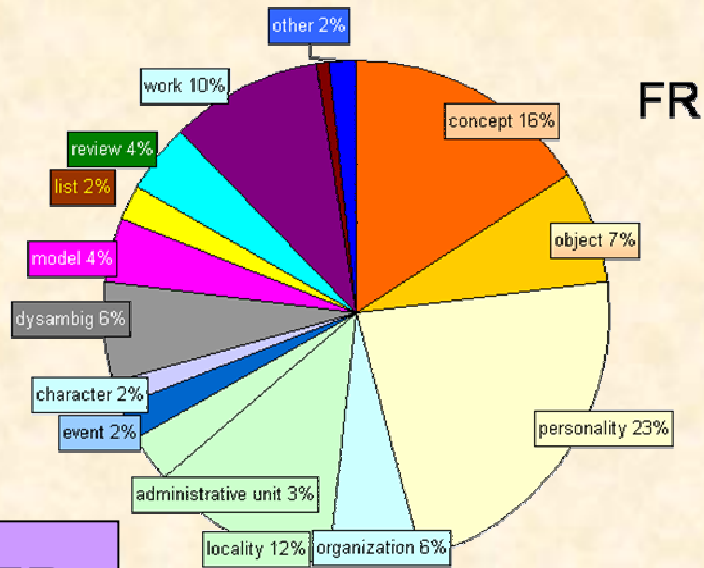
EN



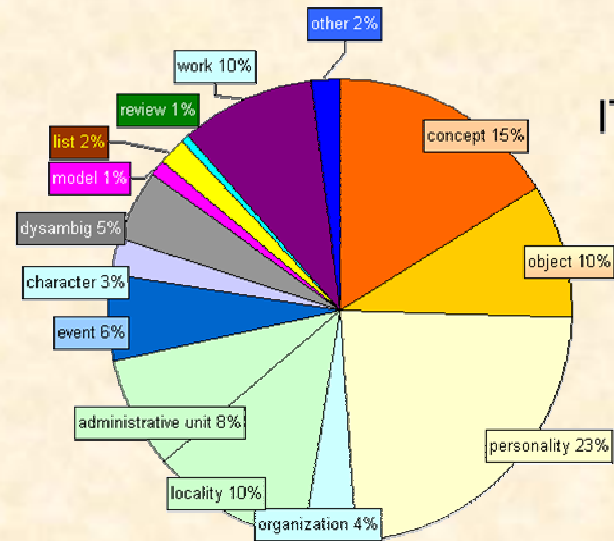
DE



FR



IT

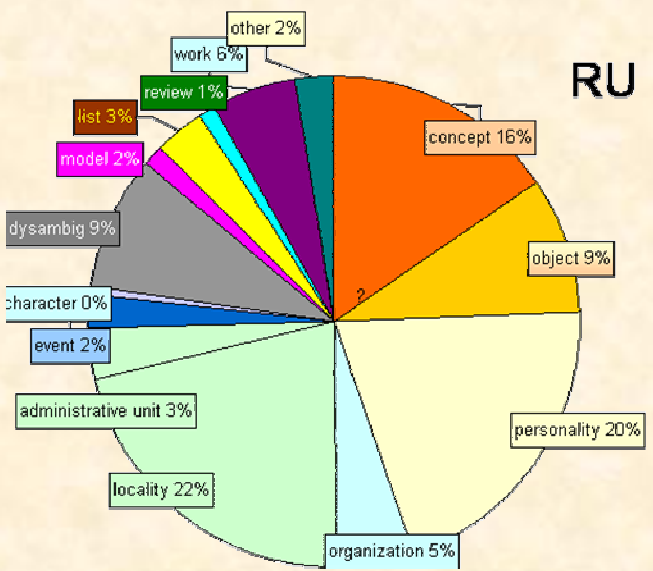
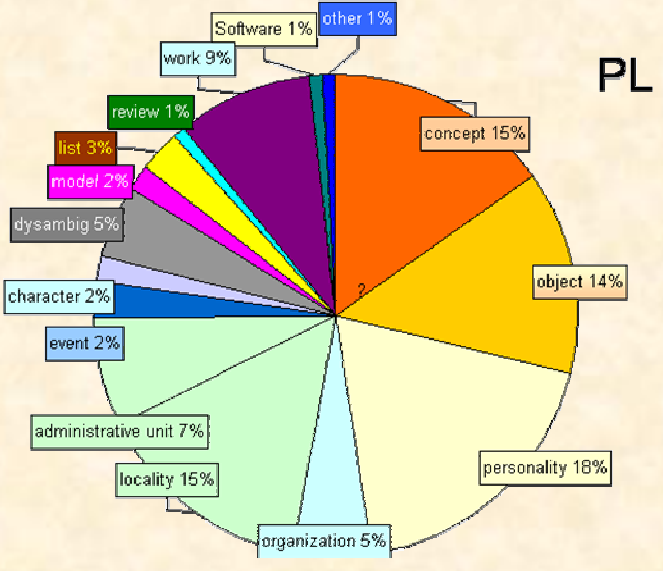


IT

PL

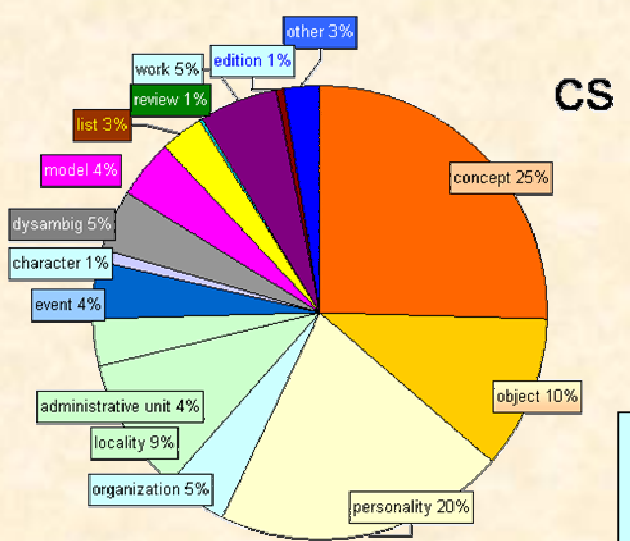
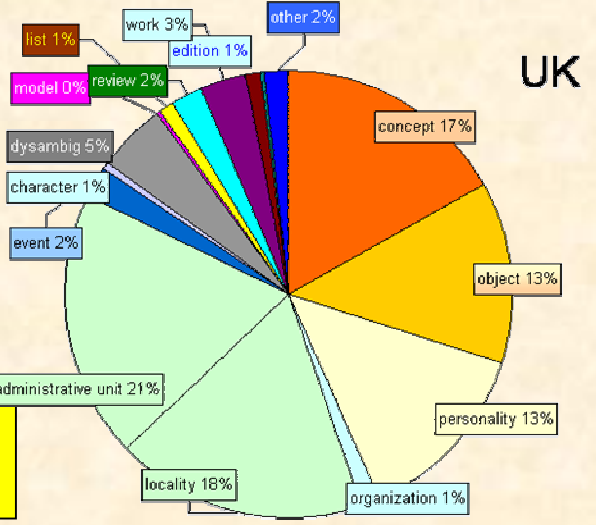
RU

4.8 Slavic Wikipedias by type



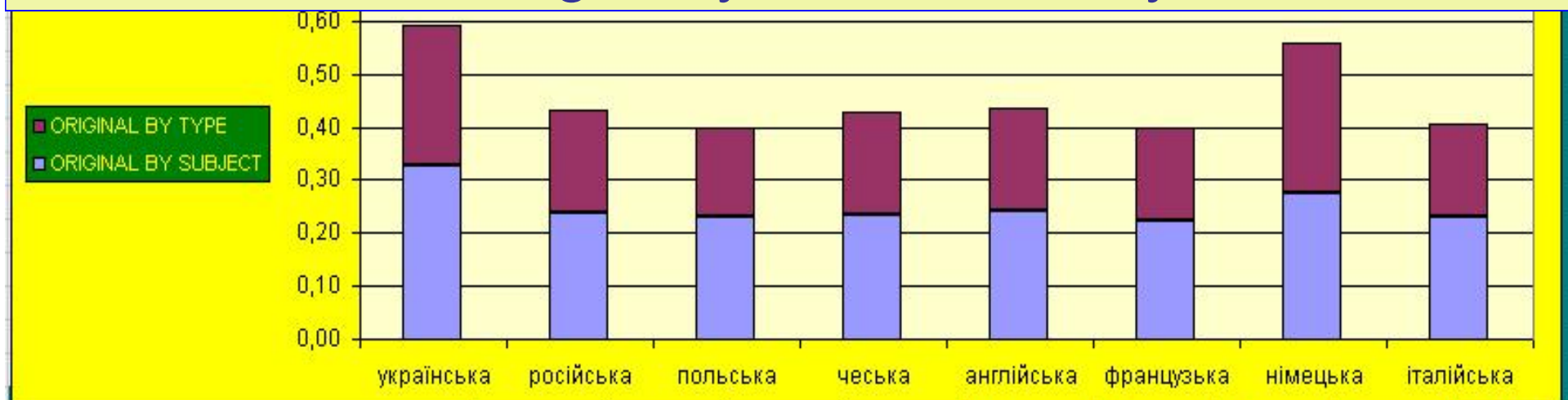
UK

CS



	українська	російська	польська	чеська	англійська	французька	німецька	італійська	середнє
ORIGINAL BY SUBJECT	0,33	0,24	0,23	0,23	0,24	0,22	0,28	0,23	0,24
ORIGINAL BY TYPE	0,26	0,19	0,17	0,19	0,19	0,18	0,28	0,17	0,20
TOGETHER	0,59	0,43	0,40	0,43	0,43	0,40	0,56	0,41	0,44

5.0 Originality or univerasality?



	UK	RU	PL	CS	EN	FR	DE	IT	разом
Ukrainian	0,00	0,53	0,51	0,68	0,80	0,72	0,88	0,62	0,59
Russian	0,53	0,00	0,33	0,50	0,51	0,44	0,68	0,46	0,43
Polish	0,51	0,33	0,00	0,42	0,50	0,41	0,64	0,37	0,40
Czech	0,68	0,50	0,42	0,00	0,42	0,35	0,63	0,41	0,43
English	0,80	0,51	0,50	0,42	0,00	0,36	0,46	0,43	0,44
French	0,72	0,43	0,40	0,34	0,35	0,00	0,56	0,36	0,40
German	0,88	0,68	0,64	0,63	0,46	0,56	0,00	0,60	0,56
Italian	0,62	0,46	0,37	0,41	0,43	0,36	0,60	0,00	0,41
together	0,59	0,43	0,40	0,43	0,43	0,40	0,56	0,41	0,45
coefficient of Slavic WP compatibility	1,74	2,20	2,38	1,86	1,79	2,09	1,41	2,15	2,16
coefficient Western WP compatibility	1,33	1,92	2,08	2,20	2,42	2,34	1,85	2,16	2,23
direction	1,31	1,14	1,14	0,85	0,74	0,89	0,76	0,99	0,97

RESUME

- Different Wikipedias have the same thematic areas, but their proportions are different.
- These differences seem to be not relevant with size of Wikipedias
- These differences seem to have some relevance with differences in culture of nations which they represente but...
- More researches required...

THANK YOU

- Made by **Andrij Bondarenko**, Wikimedia Ukraine
- More information:
 - [Користувач:А1/Статистика](#)
 - [Wikipedysta:А1/statystyka](#)
 - [User:А1/statistics](#)
- [Copyleft:cc-by-3.0](#)