Publ. Astron. Obs. Belgrade No. 80 (2006), 211 - 217 *Contributed paper*

ELECTRONIC ENCYCLOPEDIA OF ASTRONOMY AND LOCALIZATION PROBLEMS

TIJANA ZEČEVIĆ, STEVO ŠEGAN and DANILO ŠEGAN
Faculty of Mathematics, Studentski trg 16, 11000 Belgrade, Serbia and Montenegro
E–mail: gemma@matf.bg.ac.yu
E–mail: ssegan@matf.bg.ac.yu
E–mail: mm01142@alas.matf.bg.ac.yu

Abstract. The first version of en electronic encyclopedia in the Serbian language is presented. The localization problem (especially the Cyrillic letters and historical terminology) is mentioned as well. The basic tools for realizing such a project are described.

1. INTRODUCTION

There is great awareness in the last few years about the growing need in using the internet technologies, not only in the direct communication via e-mail, but in collecting informations for scientific and research works. The number of internet users is growing day by day. According to the statistics published on the Internet World Stats web site (http://www.internetworldstats.com) the internet users growth for the last 5 years in the countries of the European Union is about 141.5% and in the rest of Europe that number is 385.8%. For the last 5 years in Serbia and Montenegro the number of internet users has grown by 200%, and at this moment there is 1 200 000 internet users in our country. Data that are downloaded from this web site are shown in Tables 1 and 2 and with regard to the fact that they were updated on September 30, 2005 we can say that we have quite a new information[1].

[1]There are even newer data updated on November 9, 2005 and they can be seen on http://www.internetworldstats.com

Table 1: Internet Usage in Europe

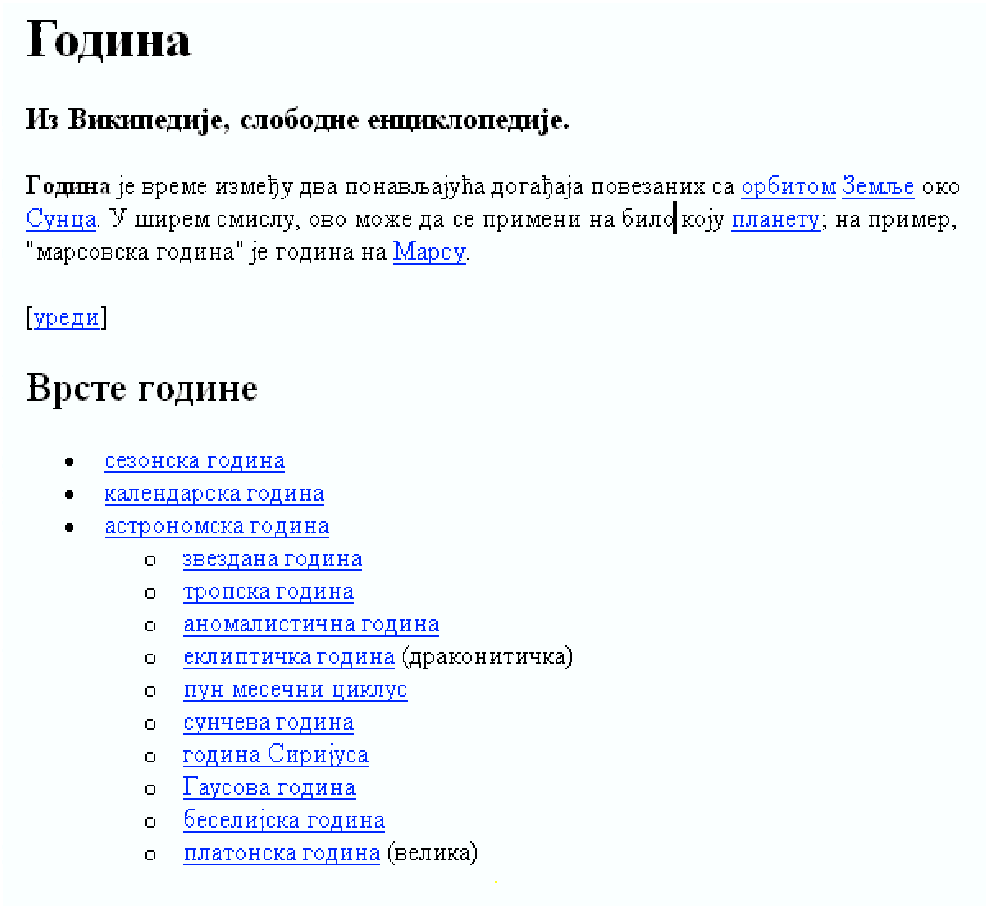
EUROPE	**Population (2005 Est.)**	**% Pop. of World**	**Internet Users (Latest Data)**	**Use Growth (2000-2005)**
European Union	460 270 935	7.20%	225 006 820	141.50%
Rest of Europe	270 747 588	4.20%	48 256 135	385.80%
TOTAL:	731 018 523	11.40%	273 262 955	165.00%
Rest of World	5 689 084 199	88.60%	684 490 717	165.40%
TOTAL:	6 420 102 722	100.00%	957 753 672	165.30%

Table 2: Internet Usage in Europe (by countries and including the EU)

EUROPE	**Population (2005 Est.)**	**Internet Users (Latest Data)**	**Use Growth (2000-2005)**
Albania	3 087 159	75 000	2 900.0%
Andorra	68 584	24 500	390.0%
Belarus	9 755 025	1 600 000	788.9%
Bosnia	9 755 025	1 600 000	788.9%
Bulgaria	7 506 098	2 200 000	411.6%
Croatia	4 459 137	1 303 000	551.5%
European Union	460 270 935	225 006 820	141.5%
Faroe Islands	49 329	25 000	733.3%
Gibraltar	26 544	6 200	287.5%
Guernsey & Alderney	63 456	20 000	0.0%
Iceland	294 947	225 000	34.3%
Jersey	88 730	8 000	0.0%
Liechtenstein	34 927	20 000	122.2%
Macedonia	2 040 389	392 671	1 208.9%
Man, Isle of	75 134	-	-
Moldova	3 902 448	406 000	1 524.0%
Monaco	33 044	16 000	128.6%
Norway	4 606 363	3 140 000	42.7%
Romania	21 377 426	4 940 000	517.5%
Russia	144 003 901	22 300 000	619.4%
San Marino	30 472	14 300	472.0%
Serbia & Montenegro	10 681 177	1 200 000	200.0%
Svalbard & Jan Mayen	2 281	-	-
Switzerland	7 452 101	4 836 671	126.6%
Ukraine	46 655 272	5 278 100	2 539.1%
Vatican City State	768	93	0.0%
TOTAL Europe	731 018 523	273 262 955	165.1%

2. ABOUT THE BAD INFORMATION AND LOCALIZATION

There is a very interesting and famous project that began in the year 2001, known as Wikipedia. Its basic idea was to make completely free multilingual encyclopedia, articles of which can be added or changed by anyone who wants to contribute. Quite a good idea if you have on mind the amount of data you can get in this way, but what about the quality of data? It is quite understandable that the open nature of Wikipedia allows some inaccuracy. Let us see the example shown in Figure 1 that is downloaded from www.wikipedia.org in the Serbian language:



Година

Из Википедије, слободне енциклопедије.

Година је време између два понављајућа догађаја повезаних са орбитом Земље око Сунца. У ширем смислу, ово може да се примени на било коју планету; на пример, "марсовска година" је година на Марсу.

[уреди]

Врсте године

- сезонска година
- календарска година
- астрономска година
  - звездана година
  - тропска година
  - аномалистична година
  - еклиптичка година (драконитичка)
  - пун месечни циклус
  - сунчева година
  - година Сиријуса
  - Гаусова година
  - беселијска година
  - платонска година (велика)

Figure 1: An example of unprecise and inaccurate data in Serbian downloaded from www.wikipedia.org

Let us suppose that I am a high school student who is interested in astronomy and that I want to find out some additional information in the area I am interested in. How many bad terms have I learnt in here? This example is quite a good illustration about the meaning of a bad information.

Beside the quality of information it is essential for information in the mother language to be accessible. The English language is dominant so it is quite natural that we can find the most articles about astronomy in English, but in the case of our country it is the same as with the printed material. So the students are enforced to collect information in other languages (mostly English) which leads us to the terminology problem. What if someone has had no opportunity to hear or read some adequate Serbian terms? In such a case a term will be translated the way translator thinks it is and it is usually wrong serbian term, as you could see in the wikipedia example.

The problem of the lack of astronomical terminology in serbian has been noted in the sixties. Despite some attempts no advances have been made in the preparation of an astronomical vocabulary. In that manner, publishing the digital astronomical encyclopedia is essential for organizing and systematizing the chaos in Serbian astronomical terminology. In fact some of Serbian astronomers covered some astronomical fields as they worked on the lexicon before, **and their work was electronically saved as data under Olivetti's PCOS operating system**. **WE HAVE REREAD** these data (in cooperation with Mr Djuro Božičković) and that will be used as **a part of comparative basis** for the future work on the digital encyclopedia.

The influence an electronic encyclopedia could have is not only a **passive** one (in the sense of preserving some historical terms), but also an **active** one, in the sense of defining and introducing new terminology. If we take into account the growth of internet users, the consequences are completely clear: spreading of introduced terminology can not be stopped. It is obvious that such a kind of project needs the presence not only of astronomers, but also of linguists.

It is true that a lot of "historical" terms and notations is used in the Serbian language, but they are not usual in the existing astronomical literature in the world. One of the tasks of the astronomical encyclopedia is to make a comparison between them to help students in easier understanding and reading both, English and Serbian literature.

There is a lot of projects in the world today that are dealing with digital preserving of thev national cultural and scientific heritage. Why digital? Because of the evident role the web has in educating people and also spreading the cultural and scientific information to the public. In that sense a working group within the MINERVA (Ministerial Network for Valorizing Activities in Digitization) project has published a handbook for quality in cultural web sites with some basic criteria that need to be adopted by the cultural and scientific entities for their web applications to become more informative and easier for searching.

The essential thing in introducing an astronomical encyclopedia in the Serbian language is the existence of its Cyrillic version. With the UNICODE standard the problem of displaying Cyrillic letters on the internet becames simpler. The Unicode assigns a unique number for every character with no regard to the platform, the software and the language. The computers work with numbers in their basis and

every letter or anything else is represented with numbers. Before the unicode there were a lot of different code pages that were often in collision. For example two different code pages gave us the same number for different characters or different numbers for the same character. Also, for most European languages to be covered it is necessary to define a smaller number of letters than it is necessary for the Latin letters in the Serbian language, not to mention the Cyrillic letters. But with the Unicode standard the Serbian Cyrillic and Latin letters are displayed correctly. This standard is supported by all modern browsers and operating systems and almost all of the other products. The Unicode assures data to be transmitted through different systems and always to be recognized. Of course, there is always the possibility for the browser not to recognize some characters, but that only means that the system is not configured correctly. There are two basic steps in the problem of recognizing the characters:

- installing the adequate fonts for corresponding characters
- configuring browser to use those fonts

Table 3 gives us an example of Cyrillic letter Ž () in different code pages

Table 3: An example of Cyrillic letter in different

Character encoding	Case	Decimal	Hexadecimal	Octal	Binary
Unicode	Capital	1046	0416	002026	0000010000010110
	Small	1078	0436	002066	0000010000110110
ISO 8859-5	Capital	182	b6	266	0010110110
	Small	214	d6	326	0011010110
KOI 8	Capital	246	f6	366	0011110110
	Small	214	d6	326	0011010110
Windows 1251	Capital	198	c6	306	0011000110
	Small	230	e6	346	0011100110

code pages

HTML representations of the same letter are Ж or Ж for capital and ж or ж for small letters.

The task of **converting** the Latin letters into the Cyrillic ones and vice versa is not so simple but as long as the unicode standard is used, it is not difficult as well, and we can say for sure that **displaying** the Cyrillic on the internet is not a problem anymore.

3. ELECTRONIC ENCYCLOPEDIA - CAN WE DO THAT AS THE REST OF THE WORLD DOES

It is always a big question how to choose the tools for realizing such a project like an electronic encyclopedia. For the decision what to use it is important to define the data

the encyclopedia should consist of. Are those only some textual data, or also many of astronomical photos should be published? Or maybe the old printed astronomical books should be digitized and presented to the public? And what about multimedia content? Why not present interesting astronomical movies within an encyclopedia?

Also, we found out that the idea that anyone can add or change some articles (as in Wikipedia) is not suitable here, but the question is if we should make a possibility for everyone to send some changes or a new article and these to be automatically forwarded to qualified persons before publishing? We could increase the content in this way and also ensure its quality. All of this defines a problem of user interaction with the encyclopedia.

It is also important to ensure a possibility for easy changing and developing an encyclopedia, itself. It is not unusual that requirements about the content change from time to time, so the flexibility in changing the structure of the encyclopedia, itself is very important.

The fact that the choice of tools used in realizing an encyclopedia is crucial, led us to FEDORA (Flexible Extensible Digital Object Repository Architecture) that can sort as some kind of database which can host a different digital content. It began as a project of the Cornell University in 1997, but its true development began in 2002 with the aim to develop a good xml and web services based system for representing different digital contents.

Digital contents are not only the documents we get by digitizing the existing non-digital objects like books, textual data, maps, pictures etc, but quite complex dynamic objects like video and audio streams. Due to their complexity it is important how to enable those documents to be represented in different formats or to be transformed during the user query. For example, if you store some old digitized book and the user wants to see only one chapter of it, how to enable this without reading the book as a whole and all of its chapters separately? Or how to enable different levels of interaction for different users? This was borne in mind through the development of FEDORA and that system gives us some advantages like:

- The objects are defined with unique identities and thus they are completely independent of the IP address of the computer
- The objects can be interconnected in different ways
- The objects that are stored can be of completely different types (books, pictures, video, audio...)
- Standard protocols for accessing the information about the objects an their content are used
- Scalability is enabled (you can put more than 10 millions of objects in FEDORA)
- Implemented security (the users can be authenticated and security policies can be implemented)
- The objects can be dynamically transformed (This means that we can store for example only one picture in high resolution, and the user can ask for only one

part of a picture, or the whole picture in smaller resolution, or zoomed picture, and he/she would succeed without us to store all of those objects separately. One object is dynamically transformed during the query and published in the desired format.)

All of these advantages make FEDORA suitable for the encyclopedia. Moreover FEDORA is freeware and it has its versions for both windows and linux operating systems. On the other hand the way in which the objects are interconnected could be changed during the development of the encyclopedia. What does that mean? If we use a standard relational database for digital content we have to know the way the objects are connected and if we want to change something in their connections we practically have to do the same job from the beginning. With FEDORA it is not the case, and it is extremely important in the beginning of developing the encyclopedia when we do not have completely the view how it will look like. Also if we use relational database and if we want to show a picture and a part of a picture we should place the two objects (picture and its part) in the database. With FEDORA it is enough to store only one object that can be dynamically transformed but it also means that it is computer intensive and we have to put that in mind when making a server for the task of encyclopedia.

There was an old typewriter printed encyclopedia of astronomy that was written by profs. Branislav Ševarlić and Jelena Milogradov-Turin in 1985, and it was misplaced somewhere at the Department of Astronomy. **It was even in the printig process but it has been never finished because of lack of money.** Prof. Stevo Šegan has found it and decided to use it as a *comparative* basis for the electronic encyclopedia. Naturally it has to be checked out but we have already begun its conversion into a digital form which should be the basis for the future work together with the data from the old Olivetti computer that were mentioned before.

It is quite clear that an encyclopedia should not be a one man project, but rather of a group of astronomers, programmers and linguists as well. It is also clear that the time needed for completing such a project is long because a lot of physical work should be done (typing the data, scanning...), then programmers' work (making web interfaces for objects manipulation, making a web site...), astronomers work (checking out the data correctness, defining an adequate terminology in the Serbian language...), as well as linguists work that can not be done by astronomers only no matter what kind of education they have.

It is obvious that at this stage, the encyclopedia is all that it should not be, but as we have chosen the tools for preparing it, as we decided about its basis and started in digitizing the data, it is surely a good beginning. But only if we become aware of a great responsibility of such a temptation, we will do it as the rest of the world does.

References

http://www.internetworldstats.com/
http://www.minervaeurope.org/
http://www.fedora.info/
http://www.wikipedia.org/