Artificial Immune System // J. Timmis and P.J. Bentley (eds.): First International Conference on Artificial Immune Systems (ICARIS'2002). – Canterbury, UK, 2002. 15. Forrest S., Perelson A., Allen L., and Cherukuri R. Self-nonself discrimination in a computer // Proc. IEEE Symp. on Research in Security and Privacy. – 1994. 16. D'haeseleer P., Forrest S. and Helman P. An immunological approach to change detection: algorithms, analysis and implications // Proceedings of the 1996 IEEE Symposium on Computer Security and Privacy. – Oakland, CA, 1996. 17. Kim J. and Bentley P. An Evaluation of Negative Selection in an Artificial Immune System for Network Intrusion Detection // GECCO 2001: Proceedings of the Genetic and Evolutionary Computation Conference. – San Francisco, California, USA. – 2001. 18. Pawlak Z. Rough Sets // International Journal of Computer and Information Science. – 1982. – № 11. 19. Bazan J.G. A comparison of dynamic and non-dynamic rough set methods for extracting laws from decision tables // Rough Set in Knowledge Discovery: Methodology and Applications. – Philadelphia: Physica-Verlag, 1998. – Chapter 17. 20. Bazan J.G., Skowron A., Synak P. Dynamic reducts as a tool for extracting laws from decision tables // Proc. of 8th International Symposium (ISMIS'94). – Charlotte (USA), 1994. 21. Wroblewski J. Finding minimal reducts using genetic algorithms // Proc. of 2nd International Joint Conference on Information Sciences (JCIS'95). – North Carolina (USA), 1995.

UDC 004.415

K. Kudim, G. Proskudina
Institute of Software Systems, National Academy of Sciences of Ukraine

COMPARISON OF EPRINTS 3.0 AND DSPACE 1.4.1 DIGITAL LIBRARY SYSTEMS

Наведено результати порівняльний аналізу основних функціональних можливостей та особливостей систем DSpace 1.4.1 та EPrints 3.1, як найпопулярніших систем побудови наукових електронних бібліотек. Особиву увагу приділено проблемам локалізації, сумісності зовнішніх форматів та зручності використання систем

The work is devoted to comparative analysis of the two most popular systems for creation of scientific digital library – DSpace 1.4.1 and EPrints 3.1. Special attention is given to problems of localization, external formats compatibility and usability.

1. Why EPrints and DSpace?

Nowadays more and more research and educational sources are created in digital form. As a result more and more institutions and organisations understand the necessity of a reliable place where such sources can be stored and easy accessible. A lot of articles, reports, experiment results, datasets, media data created by institutional division are placed on an individual hard drive or web-server of the division. Such data are often lost forever after restructurization of an institution. Furthermore, a lot of published works become available after a long period of time since they were actually finished. That's why the conception of digital library systems is so important.

There are a number of systems of such kind available commercially or with open source. And leading place here is kept by two rapidly developing open source projects named by DSpace and EPrints. They both support Open Archive Initiative [1] and among registered repositories [2] in September 2007 there were 254 archives running DSpace [3] and 237 – EPrints [4].

The last release versions for now are EPrints 3.0 and DSpace 1.4.1. The development of the systems is alive and is accompanied by an active collaboration with users.

These two systems are of the same class and have much in common. Here we will describe mostly their differences. This analysis may be interesting for those experts who choosing a digital library system for their institutions. Also it may be interesting for developers of such systems to see possible ways of improving the software.

2. About EPrints and DSpace

Both EPrints and DSpace are free, open-source, OAI-compliant, interoperable, equivalent in the functionality relevant to self-archiving, and even both written initially by the same programmer Southampton's Rob Tansley.

EPrints. With its origins in the Scholarly Communication movement, EPrints default configuration is geared to research papers but it can be adapted for other purposes and content. It was developed in the Intelligence, Agents, Multimedia Group at the Electronics and Computer Science Department of the University of Southampton in 2000. EPrints is freely distributed by GNU General Public License [5].

DSpace. Platform of digital repository DSpace was jointly developed by Hewlett-Packard Company and MIT Libraries, On November 4, 2002, the system was launched as a live service hosted by MIT Libraries and the source code made publicly available according to the terms of the BSD open source license [6], with the intention of encouraging the formation of an open source community around DSpace.

Functionality

EPrints and DSpace systems provide the basic functionality required for management digital repository solving tasks of long-term preservation and access and is intended to serve as a basis for the further development. The systems are designed to operate as a centralized, institutional service. Members of different communities deposit content directly via a Web user interface designed to make this depositing as simple as possible. The functional aspects of EPrints and DSpace can be summarized as follows [7]:

- a *data model* for basic organization of data is defined;
- *metadata* of various types are stored and indexed by the system;
- the system stores information about users of the system;
- while much of the effort is concerned with easing access to an institution's digital material, simply allowing full public access is not always acceptable. Additionally functions such as depositing and reviewing must be restricted to appropriate individuals. Hence the system includes an *authorization* function;
- the system must be able to accept incoming material, a process called *ingesting*;
- in some cases it is required that material or accompanying metadata entering the archive be checked or augmented by designated individuals. This process is called *workflow;*
- a material in the archive can be cited and accessed using information in a citation;
- end users should be able to explore and discover the contents of the repository. To this end, systems must offer *search* and *browse* functions;
- to further increase the possibilities for discovering material, metadata is exposed via the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) [8];
- it should be possible to notify end users of the system when new content of interest to them appears in the archive, rather than requiring them to repeatedly access archive to check this. Systems offer an automatic e-mail alerting service called *subscription;*
- it should be possible to handle any format from simple text documents to datasets and digital video;
- a *Web user interface* provides access to the above functionality.

Further we consider some of these functional aspects in more detail to investigate differences between EPrints and DSpace.

3. Architectural Differences

In this part we consider some aspects of architectural differences: data model, format of files, format of metadata and export/import.

3.1. Data Model

DSpace. Data storage in DSpace is organized in the way to reflect institutional structure which is using the system. Every DSpace site is divided into communities, which correspond to divisions of the organization such as department, labs, research centers or schools. A community is a highest level of DSpace content hierarchy. Communities may contain subcommunities, i.e. make up a hierarchy. Communities contain collections of logically connected stuff. Collection may be represented in more than one community. Every collection consists of items, which are the main archiving units. An element is owned by a single collection, but additionally can be represented in several others. An item is an "archival atom" consisting of grouped, related content and associated descriptions (metadata). An item's exposed metadata is indexed for browsing and searching. Items are organized into collections of logically-related material. Furthermore, items consist of bundles of bitstreams (files). The purpose of such bundles is to keep tightly related files together. There are examples for every object type of data model in Table 1 [3, 7].

Table 1

Example DSpace objects

Object Type	Example
Community	Institute of Software Systems
Subcommunity	Division of Computer Calculations
Collection	Reports; Conference publications
Item	"Comparison of EPrintss 3.0 and DSpace 1.4.1 digital library systems"
Bundle	HTML file and related image, presenting single HTML-document
Bitstream	A single HTML file; a single image file;
Bitstream Format	Microsoft Word version; JPEG encoded image format

EPrints. There is no such strict structural division as Community or Collection, which can play the important role, for example for narrowing a range of search in repositories. The idea of EPrints data model is that all records are equivalent and don't make up a hierarchy. Anyway the hierarchy is needed to browse repository because user might be unsure about what he is looking for exactly. The EPrints solution for this is views – the way to produce browsing of any type needed using metadata fields bound to items, e.g. browsing may be performed by Institutional divisions, or by Author, or more complicated by Year and then by Type and etc. So in EPrints data model it is possible to provide flexible support of hierarchical subject classifications (e.g. Library of Congress Classification) and organizational divisions.

Item, bundle of files, file data objects are similar to those of DSpace:. Similarly item is the storage entity which contains all the metadata needed to supply them for external use. The feature of EPrints is metadata that may be generated dynamically in a variety of formats out of the internal representation. One more significant difference is all types of stored sources are classified (book, article, thesis, etc.) and for each type the appropriate set of internal metadata fields matches.

Thus, item data objects are similar for both systems. They correspond to OAI-PMH which in brief can be presented as a resource-item-record. The resource can represent the traditional library object (book, article), and other entities (image, film). An item is a component of repository from which metadata about a resource can be disseminated. An item stores or dynamically generates metadata about a single resource in multiple formats, each of which can be collected in the form of record through OAI-PMH.

Hierarchical structure of items quite differs. DSpace presents more rigid system, although it covers majority of repository needs. EPrints allows to make more complex hierarchies with case on basis of external representations. In general we can conclude, that EPrints data model is more universal having its own highs and lows.

3.2. File Formats

Every file stored in the system is bound with definite format. As storage service is an essential feature of library systems so it is important to capture the specific formats of files that users submit. Integral part of file format is explicit or implicit idea of how to interpret the file content.

A list of supported file formats has much in common for both systems (see Table 2). The systems allow to store and give access to any file type [9]. However the question is if the system knows that format. There are three types of file formats in DSpace: supported, known and unsupported. It is used for automatic metadata extraction, or full-text indexing.

Table 2

Supported file formats

EPrints 3.0	DSpace 1.4.1	
HTML	Adobe PDF	MPEG
PDF	AIFF	MPEG Audio
Postscript	audio/basic	Photo CD
Plain Text	BMP	Photoshop
MS PowerPoint	FMP3	Postscript
MS Word	GIF	RealAudio
Image (JPEG)	HTML	RTF
Image (PNG)	image/png	SGML
Image (GIF)	JPEG	TeX
Image (BMP)	LateX	TeX dvi
Image (TIFF)	MARC	Text
Video (MPEG)	Mathematica	TIFF
Video (QuickTime)	Microsoft Excel	Video Quicktime
Video (AVI)	Microsoft Powerpoint	WAV
	Microsoft Project	WordPerfect
	Microsoft Visio	XML
	Microsoft Word	

3.3. Metadata Formats

DSpace. Every item in DSpace system has a record of metadata in qualified Dublin Core (DC) format [10-11]. Other metadata can be saved for the item in an attached file, but DC ensures interoperability and easy discovery of items. The DC records may be entered manually by users or they may be obtained from other metadata during submission process.

EPrints processes different record types describing documents. Every document type has its own metadata field set (subset of all the metadata fields of EPrints). The set includes only records used by that document type. Web pages are generated to present only those fields from the set of related document type.

There are following document types selected in EPrints:

– *article* in a journal, magazine, newspaper, not necessarily peer-reviewed, may be an electronic-only medium, such as an online journal or news website;
– *book* or a conference volume;

book section, or a chapter in a book;

– *monograph*, may be a technical report, project report, documentation, manual, working paper or discussion paper;
– *conference* or *workshop item* – a paper, poster, speech, lecture or presentation given at a conference, workshop or other event;
– *thesis* or *dissertation*;
– *dataset* – a bounded collection of quantitative data;
– *teaching resource* – lecture notes, exercises, exam papers or course syllabuses;
– *other* – something within the scope of the repository, but not covered by the other categories.

For repository integrity some metadata fields are mandatory thus they must be entered. Every input field has detailed help description depending on a document type.

3.4. Export/Import

EPrints. A variety of metadata sets is supported in EPrints. There is DC among them, which is declared by OAI-PMH as mandatory metadata set. For those repository items which are in public domain EPrints

exposes their metadata in DC format. If some OAI service queries other metadata formats, e.g. MODS [12], the system can provide a proper answer for that.

Data from EPrints may be exported in the following metadata formats:

- BibTeX – bibliography reference and publication management;
- OpenURL ContextObject – standard ANSI/NISO Z39.88-2004 for context-sensitive services, commonly used for full-text search [13];
- OpenURL Dissertation – the same standard specified for resources such as dissertation;
- OpenURL Journal – the same standard specified for resources such as journal;
- Dublin Core – standard ANSI/NISO Z39.85-2001 (as well as standard ISO 15836-2003) [10];
- DIDL – Digital Item Declaration Language, by means of which in MPEG-21 complex digital objects are described [12];
- EndNote – widely-distributed in a scientific community bibliographic format of citation references [15];
- HTML Citation – HTML citation format, which used for documents browsing and searching in EPrints 3.0 system;
- METS – Metadata Encoding and Transmission Standard [14];
- MODS – Metadata Object Description Schema [12];
- Reference Manager – standard for bibliographies creation [15];
- Refer – commonly used format of bibliographies[16];
- Simple Metadata (SimpleMDE) – the metadata set is subset of full possible metadata set and is used in case of a quick annotation [17];
- ASCII Citation – plain text format;
- EP3 XML – export in XML.

DSpace. Export and import functionality are developed as crosswalk plugins in DSpace. These are program modules translating DSpace object's metadata into definite external representation and vice versa. For example, from MODS metadata format to internal DSpace format and conversely. The plugins used are listed in configuration file.

The release of DSpace includes the following crosswalk plugins used with OAI-PMH:

- METS – Metadata Encoding and Transmission Standard;
- MODS – Metadata Object Description Schema;
- QDC – Qualified Dublin Core, which is the main metadata set of DSpace system [11];
- DIDL – Digital Item Declaration Language.

4. Usage differences

Here we consider user roles, submission flow and external usage.

4.1 User roles

EPrints. Initial configuration of EPrints presents four user groups defined by access rights:

- *minimal user* is able to browse repository content, subscribe to mailing lists, create saved searches;
- *depositor* has rights of minimal user, owns his private workspace to upload his items to, and is able to submit the items from his workspace for editor's review;
- *editor* has rights of depositor, and is able to accept, reject or remove items submitted by users to be placed in repository;
- *administrator* has editor's rights, and also can administer user profiles and repository items.

There exists possibility to configure access rights for each user group, e.g. restrict subjects reviewed by the editor, or skip editorial review step in case the item submissions are initiated by well-known trusted group of users. Such configuration can be done by editing related files. During submission process for each item is selected if the item will be public or permission to view will be given only to registered users, or only to editors and administrators (i.e. repository staff only).

DSpace has more advanced user access rights system which is tightly connected with used data model. Selected groups are the following: depositors, administrators, participants of submission process, subscribers, and users with permission to read non-public items. For each repository community a user group can be assigned which will have limited access to the community. For each collection a set either of single users or groups can be assigned, which will be among item submitters for the collection, have access to the content, play editor role, and administer the collection.

User can be assigned to several groups simultaneously. Appropriate rights are given to user of a particular group. Group and single user management are performed through web-interface and don't need, firstly, specific programming skills and, secondly, access to operating system where DSpace software is installed. Similarly, editing of community or collection access rights is performed through web-interface. Obviously, to access these functions one has to login with administrator rights.

In general we can say, that from roles and access rights point of view EPrints fits mostly for homogeneous repositories where unusual user rights aren't significant for variety of selected groups in separate archive parts. Such system is simple as it doesn't need configuration of access rights. As for DSpace, it has more flexible access rights system which allows different access restrictions for different repository parts. The feature of administration and submission process control through web-interface is easy to use in DSpace.

4.2. Submission

EPrints. When a new user registers in the system then a separate workspace is assigned to him, where he can upload his items. During submission a new item user should fulfill the following steps (workflow presented here is from the initial configuration):

1. Select a document type.
2. Upload files, create bundles of files if needed.
3. Input the item description.
4. Continue description – select a subject for the item.
5. Submit the item to editorial review, accepting license agreement.

At any step depositing of an item can be safely interrupted without data loss. Some fields are mandatory that means impossibility of submission for editorial review before giving them values.

DSpace. Analogously to EPrints, user has a workspace where submitted items are stored. Submission steps in DSpace are as follows:

1. Select a collection which will own an item.
2. Select options influencing on a set of fields available for input during next steps.
3. Input values of main metadata fields.
4. Input keywords to classify the item subject, and also input additional metadata fields.
5. Upload files.
6. Check if everything is right with uploaded file and its format; edit if there are mistakes.
7. Check if everything is right with all data entered earlier; edit if there are mistakes.
8. Accept license agreement.

Similarly to EPrints, interruption of submission process is safe.

In general, submission sequence is the same for both systems. EPrints has a little friendlier interface as, firstly, set of input fields isn't determined by option selection at additional step, but by marked document type; and, secondly, number of steps is less because of better grouping of input fields. However, DSpace system is more efficient. That's why submission process is performed faster in DSpace as a whole.

4.3. External Usage

DSpace has not so modern-looking interface by default as EPrints has. However there exist nice-designed repositories on base of it. The features available to external user are as follows:

– user is capable to browse an archive by communities and collections. Inside each collection and at the top of hierarchy there exists possibility to view a full list of included items ordered by Titles, Authors, Subjects or Date;

– simple search to look for an item by searching a keyword through all of the item metadata;
– Advanced search where different criteria may be selected to supply a keyword;
– Single HTML-help file and links to different places in it corresponding to currently viewed page.

EPrints current release has a nice visual web-interface by default. Any modification usually leave unchanged the following features:

– Several types of browsing through repository are presented. Default configuration includes browsing by year and by subject;
– Simple search; Extended search where plenty of metadata fields is presented to perform search by them and here a full text search is available too;
– Each input field is supplied with a question sign. Clicking on it unhide small explanations of what should go there in that field;
– RSS feed is available;
– Latest additions can be viewed.

5 Technical Differences

5.1. Installation and Prerequisites

Installation and initial configuration for both systems can be accomplished during one working day in the presence of:

1. Experience in software installation on host operation system.
2. Installation files of the e-library system itself.
3. Other essential software (see Table 3).

Table 3

Essential software

	EPrints	**DSpace**
Operating system	Unix-like	Unix-like
Web-server	Apache (including mod_perl)	Apache Tomcat or equivalence
Database Server	MySQL	PostgreSQL or Oracle
Programming language library	Perl with perl-Unicode-String, perl-XML-LibXML modules	Java, Apache Ant

EPrints. After installation of all software required for the system to function, initial configuration and installation of EPrints are made with single interactive script. As a result the first repository is created with default configuration. In addition, several cron jobs should be added: for fulltext indexing, mail sending and views generation.

DSpace. After installation of software required by DSpace, the system should be unpacked and configured by editing configuration file. Then additional cron-jobs should be added: database cleaning, indexing, mail sending and statistics gathering.

5.2. Multilingual Support

EPrints. Almost all language-dependent interface parts are collected in several separate files of simple structure. These files are divided into three groups: (1) system, non-repository-dependent phrases; (2) interface parts which may differ for different repositories, they represent the initial configuration for creation of a repository; (3) the configuration of each repository, it's a copy of default configuration at the beginning. Files of each group are stored in different folders, which contain subfolders for every language (Fig. 1).

Unfortunately cgi-script *set_lang* is not included in the last EPrints release, and language switching is only accessible with help of the script from a previous release. To make user language switching easy there should be placed appropriate links to the script in the web-interface template.

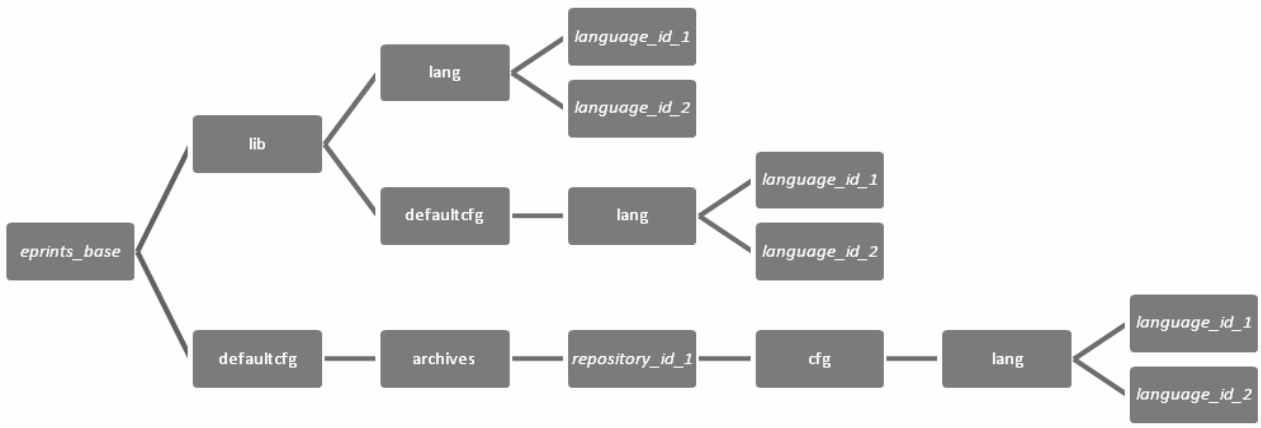

Fig. 1. Language-dependent files in EPrints

In general it's easy to make EPrints localization. And it is full enough in a sense of having full translation of language-dependent files. There are left only few phrases untranslated in interface text (for sure skipped by developers without intention to).

However there is no possibility in EPrints to present metadata of a particular repository item in different languages. It would be useful not only for web-interface views and searches, but even more for metadata harvesting. As for EPrints system there is Subject metadata field which can be translated in several languages either with subject editor tool or using related XML file. There are no ways to change structure of bibliography references in the system depending on selected language although related standards differ in different countries.

Internally EPrints uses UTF-8 encoding for all metadata fields and their values can be entered in any language, but not all of the system functionality processes this encoding correctly yet. For example, errors of such type were discovered in format of mail sent by the system, containing not only Latin letters, and in extended search function with non-Latin *Author* field.

DSpace. First of all the main system release will be described, and then patches. There is only one resource file in DSpace containing majority of interface phrases. To localize interface it's necessary to obtain translation of the file and change its name as proper. When some user enters the system through web-interface the language is determined automatically by web-browser preferences. Language switching is not accessible. E-mail message templates and help files are not included in localization. Also there is no possibility to enter news and collection descriptions in different languages.

However there are several patches for the system adding language switching, ability to translate e-mail message templates and help text. Collection and community titles, their descriptions, site news still remain untranslated.

In contrast to EPrints the metadata values can be added in several languages although not in common submission way but during optional step of editing metadata (it means additional not common operations are needed). However only the first value is used to present on screen, not depending on language. One exception is possibility to enter alternate document titles.

General conclusion is that localized interface and multilingual site can be created for both systems. In EPrints the work is easier thanks to good architectural solution. In DSpace such level of localization is achieved with use of patch. Both of the systems do not support multilingual metadata in their interfaces.

6. Conclusion

EPrints and DSpace are the systems of the same class which provide full range functionality for creation of digital repositories. They support OAI-PMH but differ in data model organization. Communities approach in DSpace is quite good, and EPrints support of various classifications has its advantages. EPrints supports more metadata formats but lacks support of QDC. EPrints is more convenient for localization, however both systems don't support multilingual metadata representation.

7. Project Reference

Analysis of series of software implementing digital library systems was carried out within the project “Design solutions for institutional automated library services” of National Academy of Sciences in Ukraine. Several systems were examined and among them were DSpace and EPrints. Both systems were successfully installed and run as live services in the local area network. Also the repository of Institute of Software Systems was created and it is publicly available at http://eprints.isofts.kiev.ua.

1. Open Archives Initiative. http://www.openarchives.org/ 2. Registry of Open Access Repositories. http://roar.eprints.org/ 3. DSpace Federation Web site http://dspace.org/ 4. GNU EPrints Software. http://software.eprints.org/ 5. Nixon W.J. DAEDALUS: Initial experiences with EPrints and DSpace at the University of Glasgow. Ariadne, October 2003. – Vol. 37. Available at http://www.ariadne.ac.uk/ issue37/nixon/intro.html. 6. Open Source BSD License. Available at http://www.opensource.org/ licenses/bsd-license.php. 7. Tansley R., Bass M., Stuve D., Branchofsky M., Chudnov D. The DSpace Institutional Digital Repository System: Current Functionality. In Proc. of JCDL 2003. 8. The Open Archives Initiative Protocol for Metadata Harvesting Protocol Version 2.0 of 2002-06-14. http://www.openarchives.org/OAI/ 2.0/openarchivesprotocol.htm. 9. A Guide to Institutional Repository Software. 3rd Edition. Open Society Institute. (2004). 10. ANSI/NISO Z39.85–-2001. The Dublin Core Metadata Element Set. – National Information Standards Organization. – 2001. http://www.techstreet. com/cgi-bin/pdf/free/335284/z39.85-2001.pdf. 11. Dublin Core Library Application Profile. http:// dublincore.org/documents/2002/09/24/libraryapplication-profile/ 12. Understanding Metadata. National Information Standards Organization. – 2004. http://www.niso.org. 13. ANSI/NISO Z39.88-2004. The OpenURL Framework forContext-Sensitive Services. National Information Standards Organization. – 2005. http://www.niso.org/standards/resources/Z39_88_2004.pdf. 14. Metadata Encoding and Transmission Standard (METS). http://www.loc.gov/standards/mets/ 15. EndNote. Bibliographies Made Easy. Getting Started Guide. Thomson. – 2006. – 86 p. http://scientific.thomson.com/media/pdfs/ ENXGettingStartedGuide.pdf. 16. Joseph F. Ossanna, Brian W. Kernighan, Gunnar Ritter. Heirloom Documentation Tools. Nroff/Troff User’s Manual. – 2007. http://heirloom.sourceforge.net/doctools/ troff.pdf. 17. Simple Metadata Annotation Specification. Version 6.2. Linguistic Data Consortium. – 2004. http://projects.ldc.upenn.edu/MDE/Guidelines/SimpleMDE_V6.2.pdf